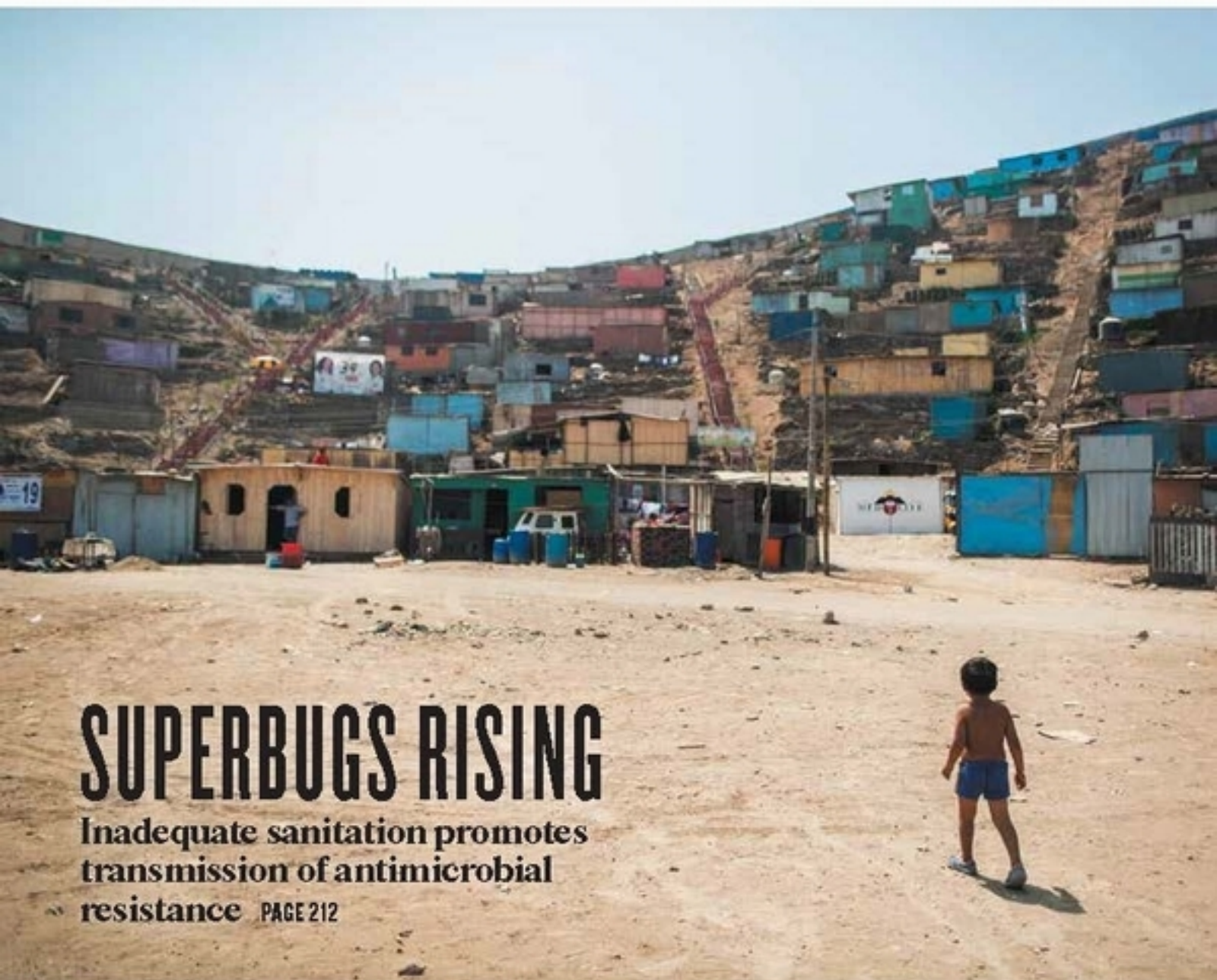


# nature

THE INTERNATIONAL WEEKLY JOURNAL OF SCIENCE



## SUPERBUGS RISING

Inadequate sanitation promotes transmission of antimicrobial resistance **PAGE 212**

### CYBERSECURITY

#### PASSWORD PSYCHOLOGY

The human fallibility factor in cybercrime

**PAGE 164**

### HUMAN EMBRYOLOGY

#### THE 14-DAY QUESTION

Are research guidelines being overtaken by events?

**PAGES 169, 182 & 251**

### ATMOSPHERIC SCIENCE

#### ECHO OF AN ANCIENT AIR

Micrometeorites record high-level Archaean oxygen

**PAGES 184 & 235**

**NATURE.COM/NATURE**

12 May 2016 £10

Vol. 533, No. 7602





# THIS WEEK

## EDITORIALS

**SMOKING** Scientists must quit arguing over e-cigarettes **p.146**

**WORLD VIEW** The compulsion to publish does real damage **p.147**



**HIGGS BISON** Online naming game spreads to Fermilab calf **p.149**

## Mothers' milk

*The safe use of medicines during breastfeeding is not an easy topic to study, but new parents deserve better information on the risks and benefits.*

When Janet Woodcock first started to practise medicine nearly 40 years ago, she quickly realized that her training had not equipped her to deal with a common dilemma. New mothers were being encouraged to breastfeed their children, but was it safe to do so if they were taking medication? “I had never received one word of information on that situation,” says Woodcock, who now heads the US Center for Drug Evaluation and Research at the Food and Drug Administration (FDA).

Woodcock's patients, she says, were “frantic” to do the best for their babies. But in the absence of data on whether and how a medicine could affect their newborn, mothers were often forced to decide between their own health and their child's. The prevailing medical advice — then and now — was, in case of doubt, to stop breastfeeding.

The situation has improved, she said at a workshop on medications and breastfeeding convened by the FDA late last month — but not nearly enough. Almost 90% of breastfeeding mothers in the United States take a medicine of some sort. For many of those drugs — including commonly used medicines to treat high cholesterol and diabetes — doctors still don't know how to counsel their patients. At the workshop, researchers illustrated how little research is done to answer those questions: a search of grants issued by the US National Institutes of Health (NIH) on the topic shows only a handful of studies, and most focus on HIV medicines.

The dearth of research comes amid renewed massive public-health pushes across the world to encourage mothers to breastfeed. Breastfeeding has been linked to fewer infections and less time in the paediatrician's office, saving parents anxiety and health systems cash. The need is particularly acute in countries where money and clean water to buy and prepare baby formula are limited. More than a decade into the twenty-first century, whether the medicines a breastfeeding mother takes are safe is a question that demands more attention.

It is undeniably difficult to conduct most clinical studies of infants. There are logistical challenges: an exhausted mother may not be keen to attend extra medical visits, and may not want to divulge the medicines she has chosen to take while breastfeeding. There are ethical challenges: clinical trials involving babies are fraught with questions about informed consent, for example. And there are financial challenges, too.

These problems have received little public attention, yet the barriers can be surmounted. At the FDA workshop, several researchers presented their success stories and lessons learned. Seemingly small measures, even changing a nappy or rocking a baby while a mother visits a clinic, can encourage women to make the effort to participate in a study. Ethical questions can be addressed through careful study design, and by paying attention to the benefits of the extra monitoring for both individual babies and for mothers. And in 2014, the FDA took a step towards raising the visibility of the matter by improving drug labels to better display what is known — and unknown — about the

safety of a given drug for breastfeeding mothers and their children.

Some researchers are already gathering data and building resources. Researchers at the University of California, San Diego, for instance, have launched the Mommy's Milk Human Milk Research Biorepository — the first of its kind, they say.

At first glance, it might not seem like sexy science for a basic researcher: the details of how particular drugs are metabolized are more the province of drug developers. And industry certainly has a responsibility to address open questions around the medicines that it produces. But basic researchers can contribute, too. Fascinating research avenues involve developmental biology, physiology and the microbiome, all of which could provide relevant information and possibly even advance fields in a fundamental way. Funders such as the NIH have taken laudable steps to address women's health issues at the level of basic research, by ensuring that animal studies include females when possible and relevant. More researchers and funders should build on that momentum and address the impact of medicines on breastfeeding mothers and their children. ■

**“Ethical questions can be addressed through careful study design.”**

## Market forces

*A European plan to commercialize quantum technologies needs a bold goal.*

Nobody ever went broke by underestimating the intelligence of the American public, goes the famous line by the US editor Henry Louis Mencken. It's actually a paraphrase, but the meaning is clear: to make money, it is safe to assume that nobody knows anything.

By rights, then, quantum physics should be extremely profitable. The subject is often used as shorthand for knowledge that is reserved for a small intellectual elite, with everyone else left scratching their heads. As Canadian Prime Minister Justin Trudeau showed last month, the quantum world is so weird that to mount even a half-decent explanation of its basic principles can bring praise and plaudits.

Can this widespread ignorance — the puzzlement at how cats can be both alive and dead, or how particles can exist in two places at once — be capitalized on? The European Commission believes that it can. Next week, it will release a plan for a continent-wide drive to turn the mysteries of quantum physics into hard cash.

This plan, called the European Quantum Manifesto, will be officially released in the Dutch town of Delft, where the commission hopes a



revolution will be born. Eyeing China, Australia, Canada and other countries that have invested huge sums of money in quantum technology, Europe does not want to miss out. With €1 billion (US\$1.1 billion) of funding, scientists and businesses will be expected to translate quantum research into quantum products to create “a more sustainable, more productive, more entrepreneurial and more secure European Union”.

These are great expectations. Europe is no doubt encouraged by the various quantum technologies that have matured in recent years. Quantum sensors, for example, can achieve high sensitivity and resolution through quantum superposition or entanglement, outperforming classical sensors in various imaging applications. Strategic use of funds could indeed take quantum sensors to market in a few years.

But for most quantum technologies, the path to commercialization is much longer and more contrived. The arguable peak of quantum technologies — the construction of a universal quantum computer — is decades, and billions of euros of targeted investment, away. But it promises perhaps the greatest gains: substantially greater power for key computations, such as simulations of chemical reactions and — maybe — machine learning.

Revolutions happen through popular uprising and not through carefully directed government investment. At some point, investors, entrepreneurs and academics are supposed to conspire on this revolution without directives from above. Hence the European Quantum Manifesto seeks to mobilize a broad base of quantum technologists. Specifically, it plans an environment in which small, high-potential quantum-tech businesses can thrive.

Given that a large majority of start-up firms fail, how is this plan supposed to work in the risky and unproven quantum-technology business? Predicting the likely outcome of the European Commission's plan is as hard as determining whether Schrödinger's cat is dead or alive without opening its box.

Can we peek inside the box to get some insights on how this commercial future might unfold? *Nature* has designed an experiment to try.

The project (see [go.nature.com/53iww6](http://go.nature.com/53iww6)) trained seven young quantum physicists to conceive and evaluate business ideas in quantum technologies. The project culminated in a presentation day last week at *Nature's* London office, where the physicists' ideas were scrutinized by a panel of experienced entrepreneurs and leaders in quantum technologies.

A PhD student from University College London invented a quantum-inspired accelerometer with a relatively safe and clear route to market.

**“This is one project that should not have to be in several places at once.”**

And two postdocs from the University of New South Wales in Sydney, Australia, have the ambition to outshine Google and IBM and build a universal quantum computer based on silicon qubits.

Two of the five ideas that were presented — an invention that permits quantum computers to be linked, and a start-up that will design quantum machine-learning algorithms — set out to depend on the few companies and groups who have already invested huge sums of money to try to build quantum-computing hardware. Both ideas are betting on being able to sell their products to only a few customers. It sounds like a risky strategy, but it might indicate a way to create and sustain the necessary critical mass of start-ups that the European Quantum Manifesto is aiming for. Focusing investment on one high-risk, high-gain goal — such as a universal quantum computer — could create a string of start-ups that each specialize in one integral component or aspect.

Still, it is unlikely that Europe's quantum-technology initiative will take this route. Given the many scientific goals in the manifesto, the authors seem to hope that the plan will have its own quantum properties and be able to address all the goals simultaneously. That looks like a mistake. It would be a missed opportunity if the quantum world that the commission hopes to create is hamstrung by the small steps and endless compromise that haunt other European projects. The initiative needs a clear and a bold goal. This is one project that should not have to be in several places at once. ■

## Smoke out

*Scientists should unite over electronic-cigarette regulation, or big tobacco will step in.*

Six million people die every year as a result of tobacco smoking, according to an estimate by the World Health Organization. It is a number worth keeping in mind as the scientific disputes over electronic cigarettes continue to smoulder.

The US Food and Drug Administration last week announced a “historic rule” that gives it the right to regulate e-cigarettes — which vaporize nicotine — as it does tobacco products. Nearly all e-cigarettes will now have to go through an approval process, with sales to young people prohibited, and health warnings included on packaging and advertisements.

Sylvia Burwell, the US Secretary of Health and Human Services, noted that e-cigarette use is shooting up among young people in the United States, “creating a new generation of Americans who are at risk of addiction”, even as cigarette smoking continues to decline.

Some states are already ahead of federal law — earlier this month, California defined e-cigarettes as tobacco products, with all that that entails. The European Union is also set to take a tougher stance. An EU-wide directive that comes into force this year on tobacco products will control nicotine content.

These ‘vaping’ devices have split researchers. Some see a route to end the tobacco scourge. Conventional medicine provides few escapes from nicotine addiction, and the speed at which smokers embrace electronic systems seems to be a blessing. If the world's smokers switched from

burning to vaping, that figure of six million deaths would fall.

But other scientists see problems. They fear that electronic devices subvert the message that smoking is bad, and offer people a nicotine fix in places where cigarettes have long been excluded. They fear a new age of nicotine, and that the six-million figure will rise.

This difference of opinion has spilled messily over into the research arena. Published studies are ruthlessly spun or picked apart by opposing sides. Sometimes the fight happens even before publication, with journalists sent quotes under embargo that critique claims and conclusions before they are publicly available.

Both sides are acting in good faith, but their arguments and increasingly entrenched positions frequently generate more heat than light. To progress, researchers on both sides must establish what evidence should be gathered to answer the central question: how can e-cigarette use and regulation lead to the largest possible reduction in deaths from tobacco? As part of this process, they should identify key data that, if forthcoming, would change their current view.

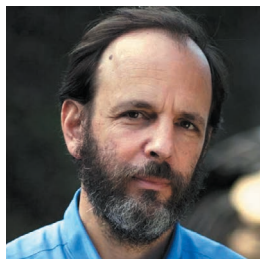
There is good reason for researchers to come together on this, and quickly. Conventional tobacco firms are grabbing an increasingly large share of the e-cigarette market. This should concern everyone — and focus minds. Few industries have historically been quite so willing to dissemble, and to market products with so few benefits and so many harms.

Researchers should remain focused on the enemy that needs to be fought — the horrific harm caused by tobacco. Disputes are part of science. They must be conducted in the open, and no researcher — and no piece of research — can be immune from criticism. But the tobacco-science community must find a way forward. It is not hyperbole to say that millions of lives are at stake. Six million of them are, every year. ■

➔ **NATURE.COM**  
To comment online,  
click on Editorials at:  
[go.nature.com/xhunqv](http://go.nature.com/xhunqv)



MICHAEL TEMCHINE



## The pressure to publish pushes down quality

*Scientists must publish less, says Daniel Sarewitz, or good research will be swamped by the ever-increasing volume of poor work.*

I am pleased to announce that as of the middle of April, my Elsevier publications had received 30,752 page views and 2,025 citations. I got these numbers in a promotional e-mail from Elsevier, and although I'm not sure what they mean, I presume that it would be even better to have even bigger numbers.

Indeed, the widespread availability of bibliometric data from sources such as Elsevier, Google Scholar and Thomson Reuters ISI makes it easy for scientists (with their employers looking over their shoulders) to obsess about their productivity and impact, and to compare their numbers with those of other scientists.

And if more is good, then the trends for science are favourable. The number of publications continues to grow exponentially; it was already approaching two million per year by 2012. More importantly, and contrary to common mythology, most papers do get cited. Indeed, more papers, from more journals, over longer periods of time, are being cited more often. One likely reason for rising citations is the incredible search capabilities that the web now affords. This would seem to be good news.

But what if more is bad? In 1963, the physicist and historian of science Derek de Solla Price looked at growth trends in the research enterprise and saw the threat of "scientific doomsday". The number of scientists and publications had been growing exponentially for 250 years, and Price realized that the trend was unsustainable. Within a couple of generations, he said, it would lead to a world in which "we should have two scientists for every man, woman, child, and dog in the population". Price was also an elitist who believed that quality could not be maintained amid such growth. He showed that scientific eminence was concentrated in a very small percentage of researchers, and that the number of leading scientists would therefore grow much more slowly than the number of merely good ones, and that would yield "an even greater preponderance of manpower able to write scientific papers, but not able to write distinguished ones".

The quality problem has reared its head in ways that Price could not have anticipated. Mainstream scientific leaders increasingly accept that large bodies of published research are unreliable. But what seems to have escaped general notice is a destructive feedback between the production of poor-quality science, the responsibility to cite previous work and the compulsion to publish.

The quality problem has been widely recognized in cancer science, in which many cell lines used for research turn out to be contaminated. For example, a breast-cancer cell line used in more than 1,000 published studies actually

turned out to have been a melanoma cell line. The average biomedical research paper gets cited between 10 and 20 times in 5 years, and as many as one-third of all cell lines used in research are thought to be contaminated, so the arithmetic is easy enough to do: by one estimate, 10,000 published papers a year cite work based on contaminated cancer cell lines. Metastasis has spread to the cancer literature.

Similar negative feedbacks occur in other areas of research. Pervasive quality problems have been exposed for rodent studies of neurological diseases, biomarkers for cancer and other diseases, and experimental psychology, amid the publication of thousands of papers.

So yes, the web makes it much more efficient to identify relevant published studies, but it also makes it that much easier to troll for supporting papers, whether or not they are any good. No wonder citation rates are going up.

That problem is likely to be worse in policy-relevant fields such as nutrition, education, epidemiology and economics, in which the science is often uncertain and the societal stakes can be high. The never-ending debates about the health effects of dietary salt, or how to structure foreign aid, or measure ecosystem services, are typical of areas in which copious peer-reviewed support can be found for whatever position one wants to take — a condition that then justifies calls for still more research.

More than 50 years ago, Price predicted that the scientific enterprise would soon have to go through a transition from exponential growth to "something radically different", unknown and potentially threatening. Today, the interrelated

problems of scientific quantity and quality are a frightening manifestation of what he foresaw. It seems extraordinarily unlikely that these problems will be resolved through the home remedies of better statistics and lab practice, as important as they may be. Rather, they would seem — and this is what Price believed — to announce that the enterprise of science is evolving towards something different and as yet only dimly seen.

Current trajectories threaten science with drowning in the noise of its own rising productivity, a future that Price described as "senility". Avoiding this destiny will, in part, require much more selective publication. Rising quality can thus emerge from declining scientific efficiency and productivity. We can start by publishing less, and less often, whatever the promotional e-mails promise us. ■

THE  
ENTERPRISE  
OF SCIENCE IS  
EVOLVING  
TOWARDS SOMETHING  
DIFFERENT AND AS  
YET ONLY  
DIMLY SEEN.

➔ **NATURE.COM**  
Discuss this article  
online at:  
[go.nature.com/sbsxfj](http://go.nature.com/sbsxfj)

**Daniel Sarewitz** is co-director of the Consortium for Science, Policy and Outcomes at Arizona State University, and is based in Washington DC.  
e-mail: [daniel.sarewitz@asu.edu](mailto:daniel.sarewitz@asu.edu)



# RESEARCH HIGHLIGHTS

Selections from the  
scientific literature

## VIROLOGY

### Zika shrinks 'mini brains' in culture

The Zika virus may trigger an immune response that causes developing brain cells to stop dividing and self-destruct.

The link between Zika infection and the birth of babies with abnormally small heads, or microcephaly, has grown stronger, but it is still not clear how the virus attacks developing brains. Tariq Rana at the University of California, San Diego, and his team grew cerebral organoids — 3D structures that model the developing brain — from human embryonic stem cells and then infected them with Zika. Over a 5-day period, uninfected organoids grew by 22.6%, whereas those exposed to Zika shrank by 16%.

Zika infection boosted the activity of a pathogen-sensing gene, *TLR3*, which has been linked to brain inflammation and degeneration. Blocking the *TLR3* protein in infected organoids lessened the damage caused by the virus.

*Cell Stem Cell* <http://doi.org/bgrx> (2016)

## BIOMATERIALS

### Second 'skin' turns back time

A polymer film that sticks to human skin reduces the appearance of wrinkles and bags under the eyes.

Robert Langer at the Massachusetts Institute of Technology in Cambridge and his colleagues designed a polysiloxane-based film that is applied to and cured on the skin. The transparent film has similar mechanical properties to skin, allowing it to conform to the surface. In small studies with human volunteers, the researchers showed that the

film reshapes the skin, making bags under the eyes look less puffy and reducing wrinkling.

The film was made from reagents that are considered to be safe for the skin. It could be used cosmetically or in wound dressings, the authors say.

*Nature Mater.* <http://dx.doi.org/10.1038/nmat4635> (2016)

## BIOPHYSICS

### Jammed microbes feel the pressure

Microbes living in a confined space can push up against each other with enough force

to physically damage their environment.

A team led by Oskar Hallatschek of the University of California, Berkeley, created a microscopic chamber that would hold roughly 100 cells of budding yeast (*Saccharomyces cerevisiae*). As the cells proliferated, they did not leave the chamber in a steady stream through a narrow exit channel but instead jammed together, building up contact pressures of almost 1 megapascal. This force was enough to cause cracks in agar gels containing growing yeast cells, and to slow down the organism's growth.

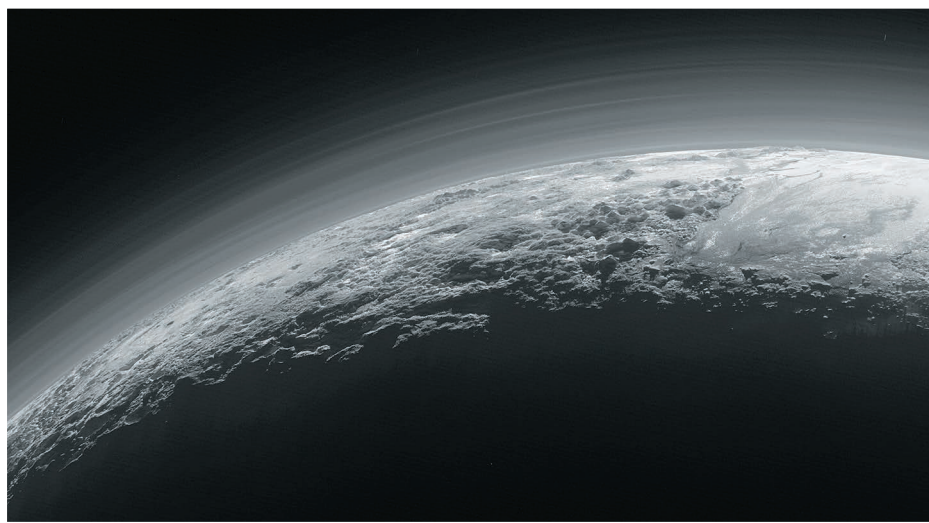
Self-driven jamming may help microbes to invade soft materials, which could contribute to biofouling — the accumulation of unwanted microbial material on surfaces, the authors say.

*Nature Phys.* <http://dx.doi.org/10.1038/nphys3741> (2016)

## CELL BIOLOGY

### Immune cell aids vascular repair

White blood cells that gobble up cellular debris also help to heal damaged blood vessels in the brain.



NASA/JHUAPL/SWRI

## PLANETARY SCIENCE

### Solar wind hits Pluto hard

The solar wind is diverted by Pluto, suggesting that, like some larger planets, the dwarf planet has a shield against the stream of energized particles emanating from the Sun.

Before NASA's New Horizons spacecraft visited the dwarf planet (pictured) in 2015, most scientists thought that Pluto interacted with the solar wind in the same way as a comet does. Comets lack protection from the wind, which diffuses around the cometary surface. But in analysing data from the spacecraft, David McComas at Princeton University in

New Jersey and his colleagues identified a 'Plutopause' — a region where Pluto's tenuous atmosphere shields the dwarf planet from the solar wind.

The Plutopause is relatively small and well defined, much like the solar-wind boundaries around Mars and Venus. Even though Pluto is small, it still exerts enough gravitational pull to keep its atmosphere sufficiently close to provide a buffer from the solar wind.

*J. Geophys. Res. Space Phys.* <http://doi.org/bgdv> (2016)



Deqin Yang at Chongqing Medical University and Lingfei Luo at Southwest University, Beibei in Chongqing, China, and their co-workers used lasers to rupture blood vessels in the brains of zebrafish. They monitored the events that followed with time-lapse microscopy, and saw the immune cells, called macrophages, migrate to the damaged area. A macrophage extended projections that adhered to the broken ends of the blood vessels and pulled them together by mechanical traction. In zebrafish engineered to lack macrophages, cerebral blood vessels healed more slowly than in normal animals.

The findings expand on the known roles of macrophages in the brain, the authors say. *Immunity* <http://doi.org/bgrv> (2016)

## MATERIALS

## Paper strips on the move

Strips of paper embedded with a conducting polymer can perform a range of movements with electrical stimulation.

George Whitesides at Harvard University in Cambridge, Massachusetts, and his colleagues made

paper actuators that expand and contract according to their water content. They added a conducting polymer that coated the fibres of the paper, and then applied Scotch tape to one side. When electrically activated, the paper heats up, dries out and contracts. When the electrical current is turned off, the paper absorbs water from the air and expands. The tape is not affected by heat or moisture, so directs the paper to bend in certain ways.

The authors made actuators of different shapes, including one that could curl up (**pictured**), and say that the devices could be used in lightweight micromachines. *Adv. Funct. Mater.* 26, 2446–2453 (2016)

## IMMUNOLOGY

## How an antibody combats HIV

A broadly neutralizing antibody against HIV can both boost people's immunity to the virus and directly target infected cells.

The antibody, 3BNC117, has previously been shown to lower HIV levels in the blood of patients. To study its effect on the immune system, Michel Nussenzweig of the Rockefeller University in New York City and his colleagues gave people with HIV one dose of the antibody. They found that patients with higher levels of the virus in their blood developed much broader neutralizing-antibody responses to HIV over six months than did those who had little to no virus (either uninfected individuals or people taking antiretroviral therapies). This indicated that the antibody is boosting the patients' ability to produce other HIV-neutralizing antibodies.

In a second study, a team led by Nussenzweig and Arup Chakraborty of the Massachusetts Institute of Technology in Cambridge showed that the same antibody speeds up the removal of

## SOCIAL SELECTION

Popular topics  
on social media

## Call calf Higgs Bison, says Twitter

Just weeks after the public voted overwhelmingly to name a new UK research vessel *Boaty McBoatface*, the US particle-physics facility Fermilab in Batavia, Illinois, asked people on Twitter to name a bison (pictured) that was born on its grassy grounds on 26 April. "What would you call our new baby bison? Tweet us with #BisonNaming. Please, no Bison McBisonface." The science world stepped up with a flood of responses — about 260 in the first day. Sandia National Laboratories in Albuquerque, New Mexico, offered "Neil deGrass Bison". Other ideas included Higgs Bison, Enrico Furry

(a play on Enrico Fermi, who discovered many radioactive isotopes), Bison Tennial and Niels Bohrson.



➔ **NATURE.COM**  
For more on  
popular papers:  
[go.nature.com/kdrhxj](http://go.nature.com/kdrhxj)

HIV-infected T cells from the blood of mice.  
*Science* <http://doi.org/bgdz>;  
<http://doi.org/bgdz> (2016)

## GENETICS

## CRISPR maps yeast genes

The CRISPR–Cas9 gene-editing system could be harnessed to speed up the search for DNA sequences linked to specific traits.

Researchers can identify genomic regions that are linked to traits, but pinpointing the responsible snippet of DNA within that region is difficult. To speed up the hunt, Meru Sadhu and his colleagues at the University of California, Los Angeles, targeted the Cas9 enzyme to cut DNA at 95 sites on one copy of chromosome 7 in yeast (*Saccharomyces cerevisiae*). The team then built a library of yeast strains, each with a genetic rearrangement at one of the 95 sites — making it easier to determine the function of a given section of DNA.

The researchers used their library to pinpoint a gene variant that makes yeast sensitive to manganese. *Science* <http://doi.org/bgd2> (2016)

## PLANETARY SCIENCE

## Planet 9 may glow from within

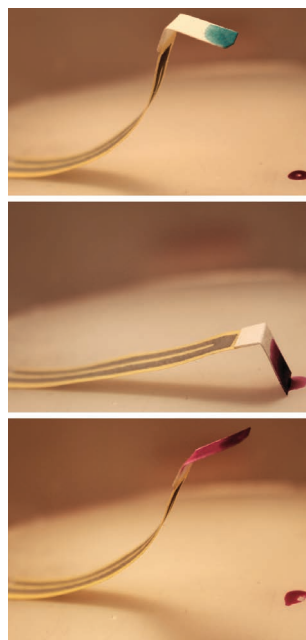
The hypothetical ninth planet of the Solar System could shine brightly.

Planet 9, if it exists, is thought to be an ice planet that is slightly smaller than Neptune, orbiting in the far outer Solar System. Esther Linder and Christoph Mordasini of the University of Bern in Switzerland modelled the evolution of the planet's probable internal structure. On the basis of its estimated mass and location, they conclude that the planet is still giving off residual heat that was generated when it was formed. This would cause the planet to emit light in the mid- and far-infrared range.

The authors say that future telescopes, such as the Large Synoptic Survey Telescope, or dedicated surveys should be able to detect Planet 9 — or rule out its existence.

*Astron. Astrophys.* 589, A134 (2016)

➔ **NATURE.COM**  
For the latest research published by  
Nature visit:  
[www.nature.com/latestresearch](http://www.nature.com/latestresearch)



# SEVEN DAYS

The news in brief

## EVENTS

### Mercury transit

NASA's Solar Dynamics Observatory captured Mercury's 7.5-hour trek across the face of the Sun on 9 May. Such planetary transits are relatively rare: Mercury and Venus pass between Earth and the Sun only occasionally. Because Mercury's orbit is tilted by 7° relative to Earth's, the innermost planet of the Solar System passes between our planet and the Sun only about 13 times a century; the next time will be in 2019. Mercury's transit of the Sun can give scientists spectroscopic clues about the chemistry of the planet's thin atmosphere.

## POLICY

### E-cig clampdown

A US clampdown on electronic cigarettes (e-cigarettes) is in the offing after the US Food and Drug Administration finalized its stance on the controversial products. The agency, which currently regulates cigarettes and rolling tobacco, announced on 5 May that it would extend its authority to cover e-cigarettes, hookah tobacco and other products containing nicotine from 8 August. This will prevent their sale to under 18s. Researchers are divided over the benefits and harms of e-cigarettes, but Sylvia Burwell, the US Secretary of Health and Human Services, said that the "drastic leap" in the use of such products was "creating a new generation of Americans who are at risk of addiction". See page 146 for more.

### TTIP opposition

The controversial trade deal being negotiated between the European Union and the United States is in jeopardy after French President

François Hollande said on 3 May that he will not accept poorly regulated free trade that questions some of France's "essential principles". His comments followed the leak on 2 May of negotiation papers concerning the Transatlantic Trade and Investment Partnership (TTIP). Among other things, critics say that TTIP could force the EU to abandon its precautionary principle in dealing with health, environmental and food-safety issues. Currently, some manufacturers and importers have to prove that their product is harmless if scientific data do not permit a complete evaluation.

Although France has no veto over the deal, it could block the negotiation process.

### Australian budget

There were no big winners or losers for science in Australia's 2016–17 federal budget, released on 3 May. But the government announced long-term investment plans for the Australian Antarctic Territory. The budget — which is released ahead of national elections in July — includes Aus\$496.2 million (US\$364 million) by 2050 to protect Australia's strategic interests in Antarctica and to sustain its scientific, environmental and economic operations there. The

government will also provide Aus\$55 million over the next decade for enhanced research and transport infrastructure in the territory, including support for over-ice convoys to research stations or drilling sites and year-round aviation access.

## PEOPLE

### UN climate chief

A seasoned Mexican diplomat and former foreign minister is to become the world's top climate official. Patricia Espinosa Cantellano, currently Mexican ambassador to Germany, was nominated on 3 May as executive secretary of the United Nations Framework Convention on Climate



OPHER SEGUIN/REUTERS

## Blazing inferno in Alberta

Almost 90,000 people were evacuated from the city of Fort McMurray last week after wildfires raged in the Canadian province of Alberta. By 8 May, the huge blaze had destroyed some 160,000 hectares of land and forest. Thanks to wet weather, fire-fighting conditions became more favourable over the weekend, but several

fires in the region are still out of control. Although the flames have not reached any of Alberta's oil-processing facilities, these have been closed, and efforts to stop the fire from spreading to neighbouring Saskatchewan may continue for weeks, authorities said. The cause of the disaster is still under investigation.



COPYRIGHT XINHUA/PHOTOSHOT  
Change by UN secretary-general Ban Ki-moon. Espinosa, a human-rights specialist with 30 years of high-level experience in international relations, is to succeed Costa Rica's Christiana Figueres, who will leave in July after six years in office.

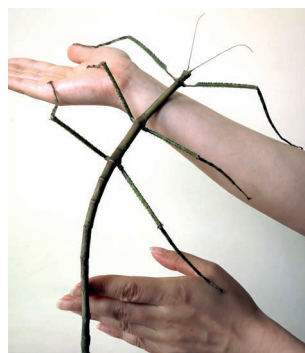
## Rockefeller head

Geneticist Richard Lifton will be the 11th president of the Rockefeller University in New York City, the university announced on 5 May. Lifton, 62, who received the 2014 Breakthrough Prize in Life Sciences for his work on the genetic causes of high blood pressure, will take office in September. Currently the Sterling Professor of Genetics and chair of genetics at Yale University in New Haven, Connecticut, Lifton will succeed Marc Tessier-Lavigne, who is leaving Rockefeller to become president of Stanford University in California.

### RESEARCH

## Giant stick insect

Scientists in China say that they have found the world's longest insect — a new species of stick insect that is well over half a metre long. Zhao Li of the Insect Museum of West China in Chengdu told the Xinhua news agency that he found the creature in 2014, on a mountain near Liuzhou City in southern



China. At 62.4 centimetres long, the insect — which has been dubbed *Phryganistria chinensis* Zhao (pictured), although it has not yet been formally described — is almost 6 cm longer than the previous record holder, *Phobaeticus chani*, a stick insect from Malaysia discovered in 2008.

## Millions for LIGO

More than 1,000 scientists and engineers will share a US\$3-million Special Breakthrough Prize in Fundamental Physics for the discovery of gravitational waves from colliding black holes. Ron Drever, Kip Thorne and Rainer Weiss, who co-founded the Laser Interferometer Gravitational-Wave Observatory (LIGO) — two US detectors where signals of elusive gravitational waves were first discovered last September — receive equal parts of \$1 million, a prize committee announced on 2 May. The

other 1,012 contributors to the experiment worldwide will share the remaining \$2 million equally. On 4 May, the LIGO team was also awarded the \$500,000 Gruber Foundation Cosmology Prize.

### FACILITIES

## Boaty no more

The United Kingdom's new polar research ship will be named after naturalist and broadcaster David Attenborough, the UK government announced on 6 May. A public vote to suggest a name for the planned £200-million (US\$289-million) vessel was overwhelmingly in favour of Boaty McBoatface. Instead, the ship will be christened RRS *Sir David Attenborough* — but Boaty McBoatface will live on as the name of the ship's remotely operated submersible.

## Chimp haven

The world's largest chimpanzee research facility is retiring its chimps. The New Iberia Research Center (NIRC) in Lafayette, Louisiana, announced on 3 May that in the summer it will begin to relocate all of its 220 primates — at up to 10 at a time — to Project Chimps, a sanctuary for chimpanzees in Blue Ridge, Georgia. NIRC was among the last labs in the world to allow invasive biomedical research on

## COMING UP

### 17–18 MAY

Details of a €1-billion initiative in quantum technology are announced at the Quantum Europe Conference in Amsterdam. See [go.nature.com/7xdf34](http://go.nature.com/7xdf34) for more.

[go.nature.com/hne6pn](http://go.nature.com/hne6pn)

### 17–22 MAY

The 10th World Biomaterials Congress takes place in Montreal, Canada.

[www.wbc2016.org](http://www.wbc2016.org)

chimpanzees before the United States effectively banned the practice in 2015 and declared captive chimpanzees to be an endangered species. NIRC says that its retirement plans had been in place since 2014.

### PUBLISHING

## Access control

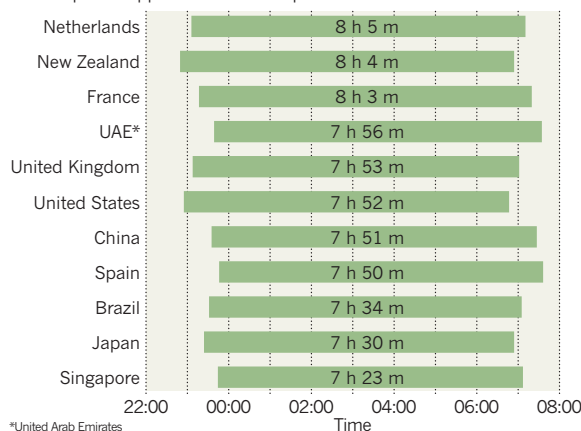
Thousands of journals are being removed from the Directory of Open Access Journals (DOAJ) in response to concerns about the increasing number of 'predatory publishers' with dubious peer-review and publishing practices. In a bid to tighten standards for inclusion, the DOAJ had asked more than 11,000 open-access journals listed on the directory to provide details about their operations. About 3,300 journals did not submit the requested information in time, so will now be delisted, says Lars Bjørnshauge, the directory's managing director. Over the past 2 years, the DOAJ has rejected more than 5,400 open-access journals, often owing to questionable publishing ethics or lack of editorial transparency. See [go.nature.com/t7aioi](http://go.nature.com/t7aioi) for more.

## TREND WATCH

Social habits in a country dictate how long its residents sleep at night. Data were collected from 5,459 people in 20 countries who used a smartphone app called ENTRAIN to record their daily bed times, wake times and exposure to outdoor and indoor light over 1 year. The analysis showed that social pressures to ignore the biological signals of sleepiness send people — particularly men — to bed late, but do not affect the time they wake up (O. J. Walch *et al. Sci. Adv.* 2, e1501705; 2016).

## GET AN EARLY NIGHT

A smartphone app reveals how sleep habits differ between countries.



# NEWS IN FOCUS

**GENETICS** Giant study of genes and education proves divisive **p.154**

**NUCLEAR SECURITY** Fears of dirty bomb create problem for biologists **p.156**

**ITALY** Row over major biomedical proposal escalates **p.158**

 **CYBERCRIME** Solving cybercrime will require behavioural science and economics **p.164**

MICHAEL BOWLES/REX/SHUTTERSTOCK



The Oculus Rift virtual-reality headset costs about US\$600.

## TECHNOLOGY

# Low-cost headsets boost virtual reality's lab appeal

*A wave of user-friendly devices is making the technology an attractive research tool.*

BY DAVIDE CASTELVECCHI

Devices that have slashed the cost of virtual reality, and transformed its performance, have implications for scientists as well as gamers. Researchers who are experimenting with the head-mounted displays say that they have the potential to find widespread use as a research tool.

Virtual reality (VR), which lets users experience a computer-generated, three-dimensional world, has produced recurring waves of hype

since the 1980s — but this time could be different, says Mel Slater, a computer scientist at the University of Barcelona in Spain who has worked in the field for two decades. Thanks to technologies originally developed for smartphones and video-gaming graphics, the performance of these headsets is now comparable to that of high-end devices that cost tens of thousands of dollars. They are sophisticated, affordable and user-friendly enough to become a staple of research labs, says Slater, rather than tools available to only very few researchers.

A gadget that has transfixed technology-news outlets is the Oculus Rift, made by Facebook-owned start-up Oculus VR of Menlo Park, California. It costs US\$600 — but operating it also requires a high-end computer that can cost more than \$1,000. Similarly priced gadgets made by smartphone-maker HTC and Sony are expected to become available this year. Vastly cheaper sets made by Google and Samsung turn a smartphone into a more basic VR device.

A lab can now buy a VR device without a dedicated equipment grant, says Anthony ►



► Steed, a computer scientist who heads a virtual-environments group at University College London.

He and Slater have been experimenting for more than a year with early prototypes of the HTC and Oculus devices, and say that the performance is just as good as that of higher-end devices, and getting better. The new devices are light enough to be worn for extended periods, and they react quickly to the user's movement, preventing the motion sickness that can occur when using VR. "Two to three years ago, the lab we used for our research cost €100,000 [US\$114,000] to set up. Now we can do the same for about €4,000," says Slater.

For years, Slater has run VR experiments with psychologists, including one that tested how white people's biases change after they have virtually inhabited the body of a black person.

Last week, Slater and Daniel Freeman, a clinical psychologist at the University of Oxford, UK, and their collaborators published a study that suggests that VR could help to treat people with severe paranoia, who often avoid crowded places because of a perception that other people want to hurt them (D. Freeman *et al.* *Br. J. Psychiatr.* <http://doi.org/bgrr>; 2016). The experimental therapy attempts to teach people

to lower their defences and to trust others by letting them visit virtual environments such as crowded lifts or underground trains.

Other studies have used VR to try to treat post-traumatic stress disorder and fear of heights or spiders. These experiments used expensive, high-end gear, but several of the researchers involved say that they now plan to start using consumer headsets instead.

As well as being cheap, the headsets are simple to set up. "It's a proper out-of-the-box experience," says Steed. If larger studies prove the therapies to be effective, patients could borrow the equipment and use it at home, Freeman says.

Neuroscientist Elizabeth Buffalo at the University of Washington in Seattle is also considering how to use the Oculus Rift. Her team studies monkeys as the animals explore interactive environments that are represented on a screen. Head-mounted sets that create a 3D environment would create a more immersive, and therefore natural, experience, she says, but current products are too big to fit on a monkey's head. "We are working on hacking the Oculus to achieve this," Buffalo says.

Creating complex virtual environments still requires specialized computer skills, says Slater. But costs are falling now that some

software developed to aid video-game companies is free to use, and many labs outsource the work. A related technology called augmented reality (AR), which superimposes images onto the user's field of view rather than replacing the scene with a different one, could also be of use in the lab, helping researchers to visualize and share data sets, says Mark Billinghurst, who studies human-computer interaction at the University of South Australia in Adelaide.

Google Glass, an early attempt at AR that projected images into the corner of a pair of glasses, was a commercial flop, but Microsoft is about to launch a more sophisticated AR headset called HoloLens. "With AR technology like HoloLens," says Billinghurst, "researchers could easily see a complex virtual data set superimposed on a real table in front of them, and also see each other face to face across the table and talk about the data."

Mary Whitton, a computer scientist who works on virtual environments at the University of North Carolina at Chapel Hill, says that there is still room for improvement in the way the systems track users' motions and in how users can interact with the virtual world using their hands. Still, she says: "I've had most fun seeing how people use what we've built in ways we never imagined." ■

## GENOMICS

# Gene variants linked to education prove divisive

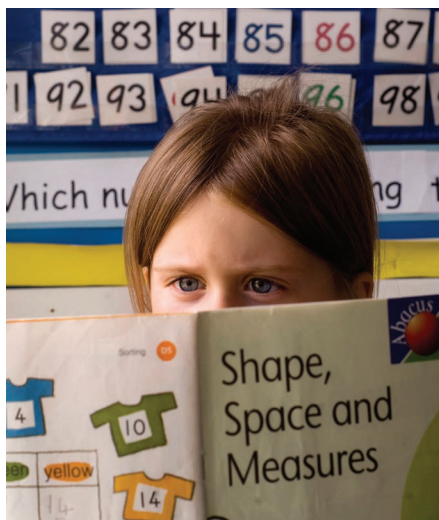
*Study uncovers 74 genetic markers that influence the number of years spent in education.*

BY ERIKA CHECK HAYDEN

The largest-ever genetics study in the social sciences has turned up dozens of DNA markers that are linked to the number of years of formal education an individual completes. The work, reported this week in *Nature*, analysed genetic material from around 300,000 people.

"This is good news," says Stephen Hsu, a theoretical physicist at Michigan State University in East Lansing, who studies the genetics of intelligence. "It shows that if you have enough statistical power you can find genetic variants that are associated with cognitive ability."

Yet the study's authors estimate that the 74 genetic markers they uncovered comprise just 0.43% of the total genetic contribution to educational achievement (A. Okbay *et al.* *Nature* <http://dx.doi.org/10.1038/nature17671>; 2016). By themselves, the markers cannot



Genetic differences explain just 3.2% of the variation in educational achievement between people.

predict a person's performance at school. And because the work examined only people of European ancestry, it is unclear whether the results apply to those with roots in other regions, such as Africa or Asia.

The findings have proved divisive. Some researchers hope that the work will aid studies of biology, medicine and social policy, but others say that the emphasis on genetics obscures factors that have a much larger impact on individual attainment, such as health, parenting and quality of schooling.

"Policymakers and funders should pull the plug on this sort of work," said anthropologist Anne Buchanan and genetic anthropologist Kenneth Weiss at Pennsylvania State University in University Park in a statement to *Nature*. "We gain little that is useful in our understanding of this sort of trait by a massively large genetic approach in normal individuals."

The study is the latest to apply genetic

TIM SMITH/PANOS

analysis to social science. Some of its authors have also studied the genetics of happiness, and plan to examine the genetics of fertility and of risk-taking behaviour.

"There's been a long-standing assumption that [genetic] differences among people are not really relevant for social-science studies," says study co-author Christopher Chabris, a cognitive psychologist at Union College in Schenectady, New York. "The main effect of this work may be the increasing realization that genetic differences matter, and now people can start to figure out how and why."

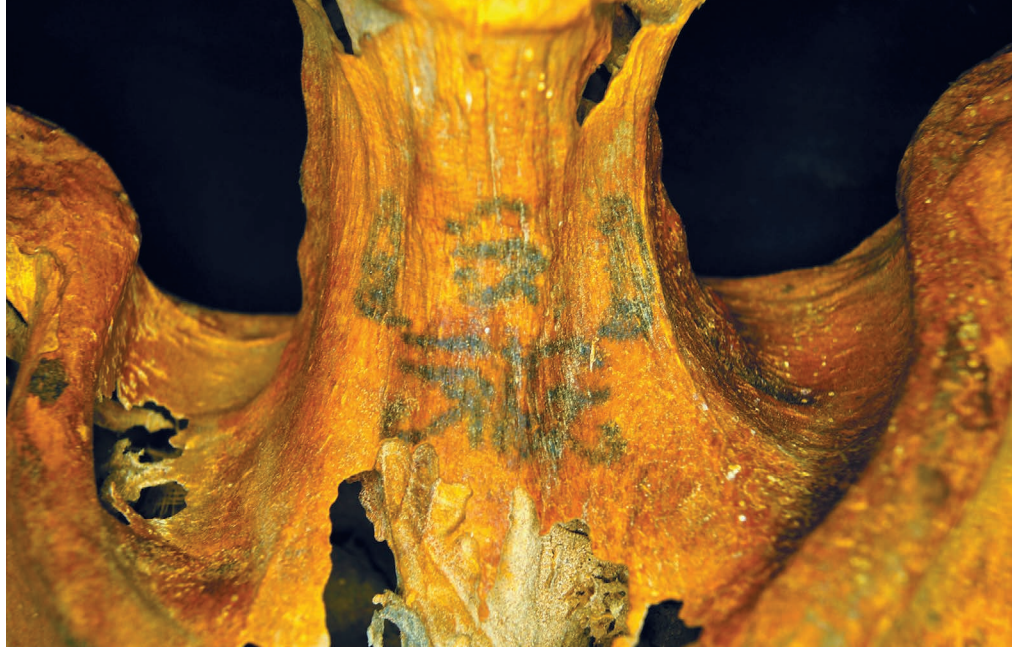
Robert Plomin, a behavioural geneticist at King's College London, agrees. The study's authors identified 9 million genetic variants that, as a group, have some influence on school success; these include the 74 genetic markers that show strong individual influence. Considered as part of an overall 'polygenic' score, the variants explain 3.2% of the differences in educational attainment between individuals. Plomin says that such studies could pave the way to predictive genetics for traits such as how well children perform on standardized tests.

Still, the researchers estimate that a person who carries two copies of the genetic variant that has the strongest known effect would complete nine more weeks of schooling over a lifetime than a person with no copies.

The authors also report that the markers they found overlap with those associated with better performance in cognitive tests, bolstering the idea that educational attainment is a proxy for intelligence. Because few large studies have tested individuals' cognitive performance, it has been difficult to discern genetic factors linked to intelligence. But it is much easier to amass large amounts of data that have sufficient statistical power to uncover genetic effects related to educational attainment, because medical studies routinely record data on participants' years of schooling.

Hsu predicts that growing knowledge of genetic contributions to intelligence could be used to help parents to select embryos created through *in vitro* fertilization. "You could allow the parents to decide whether they want to implant or not implant an embryo that has a serious cognitive impairment," Hsu says. "What is missing is the ability to know what places in the genome are affecting cognitive ability, but studies like this one will get us to that point."

But even if all the genetic contributors to educational attainment were known, the study's authors say, their effect would probably be overshadowed by other factors such as the socio-economic and educational status of a child's family. Says Chabris, "It would be irresponsible to look at a polygenic score and use it to make a prediction for a single individual." ■



The tattoos include two seated baboons depicted around a wadjet eye (top row), a symbol of protection.

#### ARCHAEOLOGY

# Sacred tattoos found on Egyptian mummy

*Unusual designs include eyes, flowers and animals.*

BY TRACI WATSON

A mummy from ancient Egypt was heavily tattooed with sacred symbols, which may have served to advertise and enhance the religious powers of the woman who received them more than 3,000 years ago.

The tattoos are the first found on a mummy from dynastic Egypt to show actual objects, among them lotus blossoms on the mummy's hips, cows on her arm and baboons on her neck. Just a few other ancient Egyptian mummies sport tattoos, and those are merely patterns of dots or dashes.

Especially prominent among the mummy's tattoos are 'wadjet eyes': possible symbols of protection against evil that adorn the mummy's neck, shoulders and back. "Any angle that you look at this woman, you see a pair of divine eyes looking back at you," says bioarchaeologist Anne Austin of Stanford University in California, who presented the findings last month at a meeting of the American Association of Physical Anthropologists in Atlanta, Georgia.

Austin noticed the tattoos while examining mummies for the French Institute of Oriental Archaeology in Cairo, which conducts research at Deir el-Medina, a village once home to the ancient artisans who worked on tombs in the nearby Valley of the Kings. Looking at a headless, armless torso dating from 1300 BC to 1070 BC, Austin noticed markings on the neck. She soon realized that they were tattoos.

Austin knew of tattoos discovered on other mummies using infrared imaging (M. Samadelli *et al.* *J. Cult. Herit.* **16**, 753–758; 2015), which peers more deeply into the skin than visible-light imaging. With help from infrared lighting and an infrared sensor, she determined that the Deir el-Medina mummy boasts more than 30 tattoos, including some on skin so darkened by the resins used in mummification that they were invisible to the eye. Austin and Cédric Gobeil, director of the French project at Deir el-Medina, digitally stretched the images to counter distortion from the mummy's shrunken skin.

The tattoos identified so far carry powerful religious significance. Many, such as the cows, are associated with the goddess Hathor, one of the most prominent deities in ancient Egypt. The symbols on the throat and arms may have been intended to give the woman a jolt of magical power as she sang or played music during rituals for Hathor.

The tattoos may also be a public expression of the woman's piety, says Emily Teeter, an Egyptologist at the University of Chicago's Oriental Institute in Illinois. "We didn't know about this sort of expression before," Teeter says, adding that she and other Egyptologists were "dumbfounded" when they heard of the finding.

Austin has already discovered three more tattooed mummies at Deir el-Medina, and hopes that modern techniques will uncover more elsewhere. ■



## NUCLEAR SECURITY

# Biologists struggle with push to end use of caesium

*Proposed switch to X-ray irradiators could affect results of research.*

BY JEFF TOLLEFSON

Anybody who wants to conduct experiments on mice in Margaret Goodell's immunology lab must submit to a host of security measures, starting with a background check by the FBI. That's because Goodell, a researcher at Baylor College of Medicine in Houston, Texas, uses a caesium-based irradiator to destroy bone marrow in mice that are set to receive stem-cell transplants. The US government fears that the radioactive caesium could be stolen to make a 'dirty' bomb.

Now the US National Nuclear Security Administration (NNSA) is working with scientists to investigate how — or whether — to replace caesium irradiators with less dangerous X-ray technology. Researchers have used the caesium devices for decades, to study everything from immunotherapy to cancer treatment, and some fear that switching to X-ray irradiators will affect their results.

Goodell, who has found subtle differences in how the mouse immune system responds to the two types of device, prefers Baylor's caesium irradiator. Her research has revealed that immune cells called B lymphocytes recovered more slowly in mice treated with an X-ray irradiator than in those exposed to caesium. But other immune cells, known as myeloid cells, rebounded faster after the X-ray treatment (B. W. Gibson *et al. Comp. Med.* 65, 165–172; 2015). Because of this, she says, "it would be difficult to compare studies using X-rays to the research that was done ten years ago".

For nuclear regulators, the risk posed by caesium is clear. The element's highly radioactive isotope caesium-137 comes in a powdered form that can be dispersed in air



Biomedical researchers often use caesium-137 to irradiate cells.

or water; exposure to the substance can cause burns, radiation sickness or death, depending on the dose. Caesium irradiators, which have long been used to eliminate pathogens in supplies of blood as well as for research applications, rely on small capsules of radioactive caesium chloride encased in a lead-covered box. There are more than 800 such devices in US medical and research facilities.

Several countries — including France, Norway and Japan — are shifting away from using caesium irradiators in blood banks because of security fears, and last year the NNSA began working with hospitals in the United States to do the same. But finding

alternative ways to treat blood is relatively simple. The NNSA is working with researchers to pin down the more complicated issue of how X-ray irradiators might differ from conventional caesium instruments for other applications.

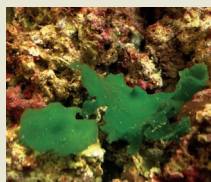
"You talk to the doctors, and they are afraid that we are going to be taking away their devices," says Maegon Barlow, director of radiological security at the NNSA. "But it's really trying to facilitate, not force."

The agency is negotiating with the Mount Sinai Health System in New York City to support a new round of studies that will compare X-ray and caesium irradiators. Jacob Kamen,

JIM R. BOUNDS/AP/PA

  
**MORE  
ONLINE**

## TOP STORY



Strange seaweed rewrites history of green plants  
[go.nature.com/jqxny](http://go.nature.com/jqxny)

## MORE NEWS

- Transparent film smooths skin back into shape [go.nature.com/qkqcj2](http://go.nature.com/qkqcj2)
- NASA jet gets a sniff of pollution over South Korea [go.nature.com/tdbohs](http://go.nature.com/tdbohs)
- Computer gleans chemical insight from lab notebook failures [go.nature.com/rl5rmx](http://go.nature.com/rl5rmx)

## NATURE PODCAST



Zika virus and birth defects; colliding quasi-particles; and psychology meets cyber-security [nature.com/nature/podcast](http://nature.com/nature/podcast)

SUZANNE FREDERICO

Mount Sinai's chief radiation-safety officer, notes that some researchers there have already conducted similar experiments.

Peter Heeger, head of organ-transplant research at Mount Sinai's Icahn School of Medicine, and his colleagues use caesium irradiators when testing immune responses in people who are going to receive organs. To predict whether a recipient's body will reject a new organ, the researchers culture B lymphocytes from the organ donor and test them against immune cells from the recipient. But B lymphocytes will not divide unless they are activated — here, by the presence of connective-tissue cells called fibroblasts. Heeger's team irradiates the fibroblasts to prevent them from replicating during this process. The scientists have run a series of unpublished experiments to determine how much X-ray radiation is necessary to suppress fibroblast growth.

"Now we know, and we are now comfortable switching for this particular procedure," says Heeger.

But Goodell says that many researchers would have to conduct lengthy experiments to ensure that they can make the transition without losing confidence in their results. Nor is she convinced that a switch to X-rays is necessary, given the security safeguards that are already in place. Anybody who needs to use the caesium irradiator at Baylor must present a security badge, enter a personal identification number and then submit to an iris scan. And if a person inside the secure room that contains the irradiator breaches any security protocols, an alarm automatically goes off in the university's security office.

"As a biologist, it's not clear to me what case has been made for [caesium irradiators] being an enormous security risk," she says.

Advocates of ending use of the devices say that the goal is to eliminate the risk of nuclear material falling into the wrong hands wherever possible. The security measures in place to protect caesium irradiators would not necessarily prevent the theft of nuclear material by somebody with permission to access these instruments, says Charles Ferguson, president of the Federation of American Scientists in Washington DC. Efforts to secure nuclear materials are often focused on this 'insider threat', as well as the disposal and recycling of irradiators, which can contain enough caesium to pose a hazard for centuries.

"I would not want humanity to lose the benefits of science," says Ferguson. "But if we can develop alternative technologies that prove comparable and can reduce the security threat to zero, I think that's a good thing." ■



## Q&A Thierry Mandon

# France's research transformation

*Thierry Mandon, who became France's research and higher-education minister last June, has vowed to cut bureaucracy in a research system that is undergoing major changes. In 2013, laws were passed to accelerate the consolidation of universities, prestigious 'grand écoles' and research-agency labs into regional clusters that could develop common research policies. And in April, Mandon announced measures to further reduce researchers' paperwork and administrative burden. He talks to Nature about what he hopes to achieve in the year remaining before France's presidential elections in 2017.*

### What are the most urgent items on your to-do list?

To simplify the rules that govern higher education, and research. To have more PhD students and researchers recruited by companies and by the public sector, and so instil a culture of research in the places where decisions are made. To help universities to develop their own sources of income, so that they can be more independent of the state. To promote a renaissance of the social sciences. And to spur the digitalization of higher education.

### What about funding? Scientists have warned that France's basic research is endangered by a lack of funding, in particular by cuts at the National Research Agency (ANR).

The ANR has seen its annual budget fall by around €250 million (US\$285 million) since 2012, to around €550 million. As a result, the success rate of grant applications is too low, at just under 10%. As President Hollande

announced in March, we plan to increase the ANR's budget by 10% this year, and by 20% next year, to bring it back to €800 million by 2018 — around the same level as its peak in 2008. We aim to boost the success rate of grant applications to between 14% and 20% in 2017.

Researchers have been critical of ANR bureaucracy. We plan to introduce a series of 50 reforms, most taking effect later this year, to lessen the administrative burden throughout the research and higher-education systems. The process of preparing grant proposals for the ANR will be greatly simplified, as will procedures for their assessment.

### What else do you hope to put in motion before next year's elections?

We want to reinvigorate the social sciences. Research in the social sciences remains very focused on publications — but value will be created through greater contact with the hard sciences and through social-science ►



► researchers informing business and policy leaders. For example, after the terrorist attacks in France last year, the CNRS [France's National Centre for Scientific Research] got all the relevant research groups working to better understand the problem of radicalization. We plan to launch new research programmes, with an initiative at the University of Paris-Sud's campus d'Orsay which will bring together much of our expertise in the social sciences.

**France has often had a reputation of lagging in innovation — yet French start-up firms seem to be an emerging force.**

France has made it easier for researchers to become entrepreneurs. But we don't succeed in making start-ups grow. Part of the problem is that the strategy of start-ups is too often to be bought out by firms in other countries. There's still too little direct contact between companies and universities, and we are working to improve this. Wealth creation must become one of the missions of the universities. Moreover, the universities are still 90% dependent on state funding. More direct links with companies could also allow universities to generate more durable financing themselves.

**Is a problem in the perception of French science that the research community seems to promote its successes much less than do, say, its US and UK counterparts?**

It is a problem. I spend my time telling French researchers to sell themselves a bit better. Take the example of the recent discovery of gravitational waves. There was a simultaneous press conference in Italy, in France and in the United States. In France, it was a low-key event in a minuscule room at CNRS, where our researchers expressed everything very modestly. By contrast, at the US event [at the National Press Club in Washington DC], one had the impression that we were at a White House event.

I'm not saying that French researchers should become as excessive as the Americans can sometimes be in their capacity to sell their advances. But in the modern world, we need to be a bit more promotional to make our excellence in research better known. At the same time, I respect a lot the sort of ethical aspects of their modesty, which has a good side.

It often seems very difficult to create change in France. But universities are innovating. My big message is that France is in the process of profoundly changing, and that researchers often aren't really taking the measure of that change. ■

#### INTERVIEW BY DECLAN BUTLER

This interview has been translated from French and edited for length and clarity. See [go.nature.com/o79dmq](http://go.nature.com/o79dmq) for a longer version.



Neuroscientist Elena Cattaneo has made complaints about the Human Technopole to the Senate.

ITALY

# Row grows over biomedical centre

*Document submitted to the Italian Senate criticizes institute that will oversee a €1.5-billion project.*

BY ALISON ABBOTT

A plan to create a €1.5-billion (US\$1.7-billion) centre for biomedical and nutritional research has been causing rifts between Italian scientists ever since Prime Minister Matteo Renzi announced it last November. Now the row has escalated, courtesy of a 48-page document submitted to the Italian Senate on 4 May by Senator Elena Cattaneo, who is also a neuroscientist at the University of Milan.

In the document, she complains that the idea for the centre, called the Human Technopole, was conceived by a small group of scientists behind closed doors, and that the large sum of money involved should not be concentrated in a single project, in particular because Italy's research community as a whole has been starved of funds for years.

"To allocate money in this way without discussion of ideas corrupts the ethics of science," Cattaneo told *Nature*.

That sentiment is in line with arguments

already made by Cattaneo and others. Cattaneo's report also lists a series of complaints against the Italian Institute of Technology (IIT) in Genoa, which Renzi has designated to oversee the Technopole project.

The complaints against both institutes are "entirely political", says Roberto Cingolani, who is the Technopole's main architect and director of the IIT. He designed the Technopole concept together with scientists from various universities and research institutes in Milan, and now plans to submit a detailed rebuttal of Cattaneo's document to the parliament.

Like the IIT, the Human Technopole was approved by government decree, and, although supported with public money, will be a private foundation. As such, it will avoid much of the red tape that holds back state universities and publicly funded research institutes.

According to Cingolani's plan, the Technopole will focus on genomics and personalized medicine, with an emphasis on nutrition,

TANIA/A3/CONTRASTO/EYEVINE

cancer and neurodegenerative diseases. The plan is now being evaluated by a panel of international scientists.

But many researchers are incensed that the project was announced without an open call for ideas. “The evaluators should have had the opportunity to compare different proposals,” says astrophysicist Giovanni Bignami, former director of the Rome-based National Institute for Astrophysics.

Earlier this year, physicist Giorgio Parisi at the Sapienza University of Rome initiated a petition, now signed by more than 72,000 people, arguing for Italy to invest more in research. But even he takes issue with the way in which the cash is to be doled out. “An investment of this magnitude should have involved the whole scientific community, and different projects should have been compared,” he says.

Supporters of the Technopole say that what matters is the progress of Italian science, not the specifics of how the project was chosen, and that the government is within its rights to set up such a centre by decree. It is “nothing unusual for a government to set science policy”, says neuroscientist Emilio Bizzi at the McGovern Institute for Brain Research at MIT in Cambridge, Massachusetts, and a member of the IIT scientific advisory board.

Cattaneo’s report also questions the choice of the IIT to coordinate the Technopole project. She notes that although the IIT is rated top among the country’s institutes for nanotechnology when measured by the impact of its publications, it is not in the top five for the life sciences or biomedicine, which are the subjects that will be the focus of the Technopole.

**“An investment of this magnitude should have involved the whole scientific community.”**

She cites a newspaper article from 6 January that reported that the IIT had not spent all of the money it received in 2013, and raised the issue of why the executive had not turned down the payments if it was not going to use them, so that they could be used by other research institutes. And Cattaneo’s report says that, according to the IIT’s internal regulations, the institute appoints members of a national committee to evaluate the institute’s progress, without the oversight of an external body.

Cingolani refutes all of these criticisms. He says that there are many ways to measure scientific success, and accuses Cattaneo of cherry-picking the facts to fit her argument. He points out that any money that the IIT doesn’t spend gets returned to the state. And he says that the IIT undergoes many levels of

evaluations and that all are carried out according to best practice. “I am preparing my rebuttal line by line, point by point,” he told *Nature*.

Parliament has yet to decide on whether to debate the issues raised by Cattaneo’s submission. But the ongoing public discussion is fuelling calls for Italy to reform how it funds research.

It is one of the few countries in the European Union without a national research agency, and in a Correspondence in this week’s *Nature*, 15 Italian members of Europe’s life-sciences organization EMBO emphasize the need for such an agency, to provide “transparent jurisdiction over the funding and execution of research” (see page 179). “The agency,” the scientists add, “would also monitor the progress of the Human Technopole and oversee its accountability.” ■

#### CORRECTIONS

The News story ‘Human embryos grown in the lab for longest time ever’ (*Nature* **533**, 15–16; 2016) wrongly characterized the US 14-day restriction on *in vitro* growth of human embryos as a law — it is a guideline. And the News Feature ‘The material code’ (*Nature* **533**, 22–25; 2016) omitted Gerbrand Ceder’s first name.



# STUCK IN THE MIDDLE

**Eric Vilain** built a career studying aspects of sex that make some people uncomfortable. Now things are getting uncomfortable for him.

BY SARA REARDON

As a medical student in Paris in the 1980s, Eric Vilain found himself pondering the differences between men and women. What causes them to develop differently, and what happens when the process goes awry? At the time, he was encountering babies that defied simple classification as a boy or girl. Born with disorders of sex development (DSDs), many had intermediate genitalia — an overlarge clitoris, an undersized penis or features of both sexes.

Then, as now, the usual practice was to operate. And the decision of whether a child would be left with male or female genitalia was often made not on scientific evidence, says Vilain, but on practicality: an oft-repeated, if insensitive, line has it that “it’s easier to dig a hole than build a pole.” Vilain found the approach disturbing. “I was fascinated and shocked by how the medical team was making decisions.”

Vilain has spent the better part of his career studying the ambiguities of sex. Now a paediatrician and geneticist at the University of California, Los Angeles (UCLA), he is one of the world’s foremost experts on the genetic determinants of DSDs. He has worked closely with intersex advocacy groups that campaign for recognition and better medical treatment — a movement that has recently gained momentum.

And in 2011, he established a major longitudinal study to track the psychological and medical well-being of hundreds of children with DSDs.

Vilain says that he doesn’t seek out controversy, but his research seems to attract it. His studies on the genetics of sexual orientation — an area that few others will touch — have attracted criticism from scientists, gay-rights activists and conservative groups alike. He is also a medical adviser for the International Olympic Committee, which about five years ago set controversial rules by which intersex individuals are allowed to compete in women’s categories.

But what has brought Vilain the most grief of late has been his stance on sex-assignment surgery for infants with DSDs. Although he generally opposes it, he won’t categorically condemn it or the doctors who perform it. As a result, many intersex advocates who object to the practice now see him as a hindrance to their cause. In November, nine bioethicists and activists resigned as advisers to his longitudinal study in protest. “I just lost my patience,” says Alice Dreger, a bioethicist who used to work at Northwestern University in Evanston, Illinois, and who was among the first to leave the study.

Although dismayed by their departure, Vilain refuses to take a stance until it is supported by

science. “The thing I don’t want to compromise is scientific integrity, even when it clashes with the community narrative.”

## BREAKING BINARY

The idea that there are only two sexes is so entrenched in society that the first question many people ask on finding out that a friend is pregnant is: boy or girl? “People don’t answer ‘I’m having a baby,’” says Vilain. “They probably should.”

At Necker University Hospital for Sick Children in Paris in the 1980s, he says, doctors presumed that a child would be psychologically damaged if he or she did not have normal-looking genitalia. In Vilain’s experience, that belief was so strong that doctors would take genital abnormalities into account when deciding how

DAVID WALTER BANKS



hard to fight to save a premature baby. “The unanimous feeling was that boys with a micro-penis could never achieve a normal life — that they were doomed,” he says. (The paediatric-surgery department at Necker refused to answer questions relating to past or current standards of care.)

DSDs occur in an estimated 1–2% of live births, and hundreds of genital surgeries are performed on infants around the world every year<sup>1</sup>. But there are no estimates as to how often a child’s surgically assigned sex ends up different from the gender they come to identify with.

What do exist, however, are stories of people who say that they have been harmed: children who struggle to fit in with peers, adolescents who are stressed, harassed or attempt suicide, and adults who are furious that they were not

involved in the decision to modify their bodies. Over the past two decades, and especially in the past few years, intersex activists worldwide — some of whom do not identify as either gender — have begun to speak out against the practice. Unless a child’s life is in danger, they argue, he or she should have the right to decide on surgery when older.

Vilain’s fascination with the biological complexities of sexual differentiation made him want to study the causes of DSDs. So in 1990, he joined the lab of geneticist Marc Fellous at the Pasteur Institute in Paris. Fellous was studying a newly discovered gene called *SRY*, which resides on the Y chromosome and is crucial in triggering the development of male features. Vilain helped to identify the causes of several DSDs, such as XY people who look female because

of mutations that disable the *SRY* gene<sup>2</sup>, and people who carry a copy of *SRY* even if they do not have a Y chromosome<sup>3</sup>. Vilain was an unusual student, Fellous says, because his clinical background allowed him to bridge lab work and patient care. Fellous says that it is often difficult to explain to the families of children with DSDs why the research would be helpful. “Eric was useful for this,” he says. “He was a very open mind, really close to families.”

In 1995, Vilain left France for a faculty job at UCLA. There, he began tackling questions about sexual development from every possible angle. He created mouse models with mutations in *SRY* or other sex-linked genes to study how their developing brains respond to hormones — research that could lead to better care for people with DSDs.



Perhaps most notoriously, he has explored the roots of sexual orientation, work that made even his colleagues uncomfortable. In 2006, he was looking to publish work by his postdoc Sven Bocklandt, who had found links between the way genes are expressed from a mother's X chromosomes and the chances of her having a gay son. When he approached biostatisticians for help, several refused to collaborate, Vilain says, because they were afraid of how the public might respond.

Studies on the genetic underpinnings of homosexuality are controversial. Religious conservatives who believe that being gay is a choice argue that scientists are trying to legitimize it; gay activists worry that the research will lead to misguided attempts to 'cure' gay people. Vilain gets occasional attacks from both groups. But he says that his colleagues' squeamishness around controversial research was unscientific. So, he stormed into the office of the UCLA biostatistics chair, Kenneth Lange, to complain.

"Eric's not afraid to kick up some dust and stand up for the people in his lab," Bocklandt says. "I think that's why he's been so successful." A statistician eventually volunteered to help.

Dean Hamer, a retired geneticist formerly at the US National Cancer Institute in Bethesda, Maryland, trained Bocklandt and has studied the genetics of sexual orientation. He says that Vilain is pretty much the only geneticist who still does serious research on the topic. "That takes a level of courage and belief that ultimately the biology will win out," he says.

#### COURTING ADVOCACY

Vilain's research and interest in policy has put him on the front lines of the lesbian, gay, bisexual, transgender and queer (LGBTQ) rights movement and has made his lab a magnet for LGBTQ students. His work also made him a sort of scientist-laureate for the intersex advocacy community, which started gaining prominence in the early 1990s with the formation of the Intersex Society of North America in Rohnert Park, California. The group, founded by activist Bo Laurent, lobbied for recognition of intersex as a human condition rather than an affliction, and opposed infant surgery.

Vilain, who met Laurent in 1997, says that she helped to shape his opinions on surgery and other topics that are important to intersex people, such as the stigma they face. Although Laurent and her colleagues were well informed and knowledgeable about the science of DSDs, they struggled to be heard in scientific conversations. "I think the view was that they were zealots," Vilain recalls.

In 2005, several paediatric societies met in Chicago to draft a consensus statement on the management of intersex conditions — a still-influential document<sup>4</sup> that guides the standard

of care. Laurent attended the meeting hoping to see the word hermaphrodite struck from the medical vocabulary. The term was not only offensive — it labelled a person rather than a disorder — it was also scientifically inaccurate because it suggested that the person had functioning male and female organs.

Rather than being heard, Laurent recalls being sidelined. But Vilain, who headed the genetics working group, met with her in secret throughout the meeting, drafting a case to present to the group. They met stiff opposition from medical doctors, who saw no reason for change, but their language was ultimately adopted in the final statement<sup>4</sup>.

Over the years, Vilain continued to build a reputation as an ally to intersex people. In 2011, when he and psychologist David Sandberg at the University of Michigan in Ann Arbor began the ten-institution registry to track children with DSDs, ethicists and activists enthusiastically joined its advisory board. Funded by the US National Institutes of Health, the Disorders of Sex Development Translational Research Network has enrolled more than 300 children, collecting medical records and blood samples and performing interviews to answer a variety of biological and psychological questions.

Many of the advocates who joined as advisers had hoped that development of the network would lead to a denouncement of infant genital surgery by revealing the damage that it can cause. "No one has demonstrated anything but harm," says Anne Tamar-Mattis, legal director of the intersex advocacy group interACT in San Francisco, California. "Research that settles that question is useful."

**"You're basically calling doctors torturers when they're doing something considered standard medical practice."**

But the study has yet to do what advocates hoped. Sandberg, who heads the network's psychological research, has collected evidence that emotional and social support from the family is the most important contributor to the psychological and mental health of a child with a DSD. He suspects that it has an even greater impact than surgery. "I never question people's experiences," Sandberg says of the activists who believe that surgery is always harmful. "What I do question is whether they're generalizable."

One argument in favour of infant surgery is that a child could be psychologically scarred by growing up with intermediate genitalia, but there is little evidence for or against that. In rare

cases, surgery could help to prevent cancer. Complete androgen insensitivity syndrome, for instance, confers an increased risk of testicular cancer that can be lowered through surgery<sup>5</sup>. But Vilain points out that the risk before puberty is very small<sup>6</sup>, suggesting that surgery could wait.

Although few surgeons were willing to talk openly about infant genital surgery, some do argue that the fear of harm is overblown or at least outdated. Laurence Baskin, a paediatric urologist at the University of California, San Francisco, says that the days of "assigning gender" are long gone, because scientists no longer believe that a child can be made to be a boy or a girl. Most DSDs can be diagnosed and the outcomes predicted; physicians use the diagnosis to advise parents on which gender the child is likely to identify with, he says. For instance, the most common cause for a DSD is congenital adrenal hyperplasia — which can result in ambiguous genitalia for XX children. Between 90% and 95% of people with the condition identify as female<sup>7</sup>.

When asked about children with this disorder who ultimately do not identify as female, another paediatric urologist — who wished not to be named — argues that the process can be reversed. People have sex-change surgery as adults all the time, he says.

Such arguments infuriate Tamar-Mattis. "If one time in 20 you're cutting a little boy's penis off, is that a risk worth taking?" she says.

Vilain doesn't think so, and doesn't generally recommend surgery to his patients. He says that in his experience, more parents are now choosing to delay surgery.

But he and his collaborators on the longitudinal study are reluctant to condemn surgery outright — they prefer to approach each case individually and to consider the views of parents who may feel strongly about what is right for their child.

This attitude helped to create the rift between the researchers and intersex advocates. At the end of 2015, Dreger, who had served as the bioethicist for the longitudinal study announced her resignation in a blogpost. "I can't continue to help develop 'conversations' around 'shared decision making' that allow decisions to be made that I believe violate the most basic rights of these children," she wrote. "I am fed up with being asked to be a sort of absolving priest of the medical establishment."

Vilain was blind-sided by the post. "I was very saddened by this," he says. "She's a friend." After her departure, eight other advocates sent the study's leaders a letter of resignation.

Most would not comment on the record, but say that they were upset that the researchers were making decisions about which questions to pursue without sufficiently consulting them.



The controversy surrounding female runner Caster Semenya's participation in international competitions helped lead to rule changes about who can compete in women's events.

OLIVER MORIN/AFP/GETTY

For example, advocates are also concerned about the psychological impacts on children from having their genitals photographed for the purpose of diagnosis or to plan treatment. Some say that Vilain became hostile in meetings. They accuse him and Sandberg of putting research interests ahead of human suffering.

"We live in a community of people who have experienced the harms of these practices," says Arlene Baratz, a radiologist who serves as medical adviser to a DSD support group and is one of those who resigned from the study. She and others say that in their decades of work as advocates they have never been contacted by someone who was helped by surgery.

Vilain says that he does talk to such patients in his practice, but because they are living happy lives, they have no reason to speak out. Without data on outcomes, says Douglas Diekema, a medical ethicist at the Seattle Children's Research Hospital Institute in Washington, it is impossible to weigh up whether surgery is overall harmful, helpful or neutral for most people. "Good ethics requires good data," he says.

But a legal battle in the United States could change medical practice before those data are in. Tamar-Mattis is one of the lawyers representing the family of a baby who underwent feminizing surgery at 16 months old. The child, now 11 years old, identifies as male, and his lawyers argue that South Carolina's Department of Social Services and the university that performed the surgery violated the child's rights.

Intersex advocates are watching the case with great interest, because it could lay the groundwork for future suits that could effectively

outlaw the procedures in the United States.

In January, the United Nations released a report saying that sex-assignment surgeries on infants "lead to severe and life-long physical and mental pain and suffering and can amount to torture and ill-treatment". Vilain and Sandberg worry that the language could alienate doctors and parents alike. "You're basically calling doctors torturers when they're doing something considered standard medical practice," Vilain says. He points out that few medical procedures are governed by law — physicians tend to operate according to guidelines and principles. "I'm not opposed to guidelines, I'm opposed to things that completely alter medical practice in an irreversible way," he says. He and Sandberg also worry that legal bans could drive infant surgery underground. "Parents are scared. You just don't dictate to them and say get over it," Sandberg says.

#### TESTING PATIENCE

Vilain's expertise has plunged him into other controversies. One example is his involvement with the International Olympic Committee, which in 2011 revised its policy on athletes who identify as female but who have male sex organs or produce high levels of testosterone.

The issue came to the fore in 2009 after 18-year-old South African runner Caster Semenya, who identifies as female, was subjected to humiliating sex testing before being allowed to continue competing in the women's category.

To head off future problems, the medical advisory board, under Vilain's leadership, drew

a bright line for the 2012 Olympics. People with testosterone levels above 10 nanomoles per litre of blood could not participate in women's events, no matter how they identify. Exceptions are made only if the athletes can prove that they are resistant to the effects of testosterone.

Many activists and ethicists are furious about the policy. "It bears noting that athletes never begin on a fair playing field; if they were not exceptional in one regard or another, they would not have made it to a prestigious international athletic stage," wrote bioethicist Katrina Karkazis from Stanford University in California in a 2012 article<sup>8</sup> lambasting the policy.

Even Vilain struggles to defend it on scientific grounds. Although women with DSDs that result in high testosterone levels are over-represented among Olympians, the hormone does not seem to directly impact their performance. "It is very imperfect," he admits. "But if we don't have a dividing line, then there is no point in segregating sexes in sports." (The policy has been temporarily suspended and is under review.)

Some of Vilain's detractors question how he can support a somewhat arbitrary call in this situation while requiring more evidence to condemn infant surgery. But sport, he argues, depends on rules and policies, whereas medicine relies on best-practice guidelines — and that is what he hopes to develop through research.

He and his collaborators plan to continue the longitudinal study. The team has recruited a bioethicist, John Lantos of Children's Mercy Hospital in Kansas City, to replace Dreger, and it still has some patient advocates involved. Vilain says that he is trying not to antagonize anyone — the next iteration will include research on more questions that the participants say are priorities, such as how to preserve fertility for people with DSDs and identifying cancer risks.

Yet Vilain's experiences with patient advocates have hardened him somewhat. "I call the ones who work with us advocates; those against us activists," he says. He remains driven by questions about sex, even if it kicks up dust. "We're trying to listen to the community, but by the same token we're committed to producing data and evidence." ■

Sara Reardon writes for *Nature* from Washington DC.

1. Blackless, M. *et al.* *Am. J. Hum. Biol.* **12**, 151–166, (2000).
2. Vilain, E., Jaubert, F., Fellous, M. & McElreavey, K. *Differentiation* **52**, 151–159 (1993).
3. Abbas, N. *et al.* *C. R. Acad. Sci. III* **316**, 375–383 (1993).
4. Lee, P. A., Houk, C. P., Ahmed, S. F. & Hughes, I. A. *Pediatrics* **118**, e488–e500 (2006).
5. Coombs, M. *et al.* *Endocr. Dev.* **27**, 185–196 (2014).
6. Deans, R., Creighton, S. M., Liao, L. M. & Conway, G. S. *Clin. Endocrinol.* **76**, 894–898 (2012).
7. Dessens, A. B., Slijper, F. M. & Drop, S. L. *Arch. Sex Behav.* **34**, 389–397 (2005).
8. Karkazis, K., Jordan-Young, R., Davis, G. & Camporesi, S. *et al.* *Am. J. Bioeth.* **12**, 3–16 (2012).





# *The human side of* **CYBERCRIME**

*As cyberattacks grow ever more sophisticated, those who defend against them are embracing behavioural science and economics to understand both the perpetrators and their victims.*

BY M. MITCHELL WALDROP

Say what you will about cybercriminals, says Angela Sasse, “their victims rave about the customer service”.

Sasse is talking about ransomware: an extortion scheme in which hackers encrypt the data on a user’s computer, then demand money for the digital key to unlock them. Victims get detailed, easy-to-follow instructions for the payment process (all major credit cards accepted), and how to use the key. If they run into technical difficulties, there are 24/7 call centres.

“It’s better support than they get from their own Internet service providers,” says Sasse, a psychologist and computer scientist at University College London who heads the Research Institute in Science of Cyber Security. That, she adds, is today’s cybersecurity challenge in a nutshell: “The attackers are so far ahead of the defenders, it worries me quite a lot.”

Long gone are the days when computer hacking was the domain of thrill-seeking teenagers and college

students: since the mid-2000s, cyberattacks have become dramatically more sophisticated. Today, shadowy, state-sponsored groups launch exploits such as the 2014 hack of Sony Pictures Entertainment and the 2015 theft of millions of records from the US Office of Personnel Management, allegedly sponsored by North Korea and China, respectively. ‘Hactivist’ groups such as Anonymous carry out ideologically driven attacks on high-profile terrorists and celebrities. And a vast criminal underground traffics in everything from counterfeit Viagra to corporate espionage. By one estimate, cybercrime costs the global economy between US\$375 billion and \$575 billion each year<sup>1</sup>.

Increasingly, researchers and security experts are realizing that they cannot meet this challenge just by building higher and stronger digital walls around everything. They have to look inside the walls, where human errors, such as choosing a weak password or clicking on a dodgy e-mail, are implicated in nearly one-quarter of all cybersecurity failures<sup>2</sup>. They also have to look outwards, tracing the underground economy that supports the hackers and finding weak points that are vulnerable to counterattack.

“We’ve had too many computer scientists looking at cybersecurity, and not enough psychologists, economists and human-factors people,” says Douglas Maughan, head of cybersecurity research at the US Department of Homeland Security.

That is changing — fast. Maughan’s agency and other US research funders have been increasing their spending on the human side of cybersecurity for the past five years or so. In February, as part of his fiscal-year 2017 budget request to Congress, US President Barack Obama proposed to spend more than \$19 billion on federal cybersecurity funding — a 35% increase over the previous year — and included a research and development plan that, for the first time, makes human-factors research an explicit priority.

The same sort of thinking is taking root in other countries. In the United Kingdom, Sasse’s institute has a multiyear, £3.8-million (US\$5.5-million) grant from the UK government to study cybersecurity in businesses, governments and other organizations. Work from the social sciences is providing an unprecedented view of how cybercriminals organize their businesses — as well as better ways to help users to choose an uncrackable yet memorable password.

The fixes are not easy, says Sasse, but they’re not impossible. “We’ve actually got good science on what does and doesn’t work in changing habits,” she says. “Applying those ideas to cybersecurity is the frontier.”

### KNOW YOUR AUDIENCE

Imagine that it is the peak of a harried work day, and a legitimate-looking e-mail lands in your inbox: the company’s computer team has detected a security breach, it says, and everyone needs to run an immediate background scan for viruses on their machines. “There’s a tendency to just click ‘accept’ without reading,” says Adam Joinson, a social psychologist who studies online behaviour at the University of Bath, UK. Yet the e-mail is a fake — and that hasty, exasperated click sends malware coursing through the company network to steal passwords and other data, and to convert everyone’s computers into a zombie ‘botnet’ that fires off more spam.

The attackers, it seems, have a much better grasp on user psychology than have the institutions meant to defend them. In the scenario above, the success of the attack relies on people’s instinctive deference to authority and their lowered capacity for scepticism when they’re busy and

distracted. Companies, by contrast, tend to impose security rules that are disastrously out of sync with how people work. Take the ubiquitous password, by far the simplest and most common way for computer users to prove their identity<sup>3</sup>. One study<sup>4</sup>, released in 2014 by Sasse and others, found that employees of the US National Institute of Standards and Technology (NIST), headquartered in Gaithersburg, Maryland, averaged 23 ‘authentication events’ per day — including repeated logins to their own computers, which locked them out after 15 minutes of inactivity.

Such demands represent a substantial drain on employees’ time and mental energy — especially for those who try to follow the standard password guidelines. These insist that people use a different password for each application; avoid writing passwords down; change them regularly; and always use a hard-to-guess mix of symbols,

**CYBERCRIME COSTS THE GLOBAL ECONOMY BETWEEN \$375 BILLION AND \$575 BILLION EACH YEAR.**

numbers and uppercase and lowercase letters.

So people resort to subversion. In another systematic study of password use in the real world<sup>5</sup>, Sasse and her colleagues documented the ways in which workers at a large multinational organization side-stepped the official security requirements without (they hoped) being totally reckless. The employees’ methods — writing down a list of passwords, for example, or transferring files between computers using unencrypted flash drives — would be familiar in most offices, but essentially created a system of ‘shadow security’ that kept the work flowing. “Most people’s goal is not to be secure, but to get the job done,” says Ben Laurie, who studies security compliance at Google Research in London. “And if they have to jump through too many hoops, they will say, ‘To hell with it.’”

Researchers have uncovered multiple ways to ease this impasse between workers and security managers. Lorrie Cranor directs the CyLab Usable Privacy and Security Laboratory at Carnegie Mellon University in Pittsburgh, Pennsylvania — one of several groups worldwide that are looking at ways to make password policies more human-compatible.

“We got started on this six or seven years ago, when Carnegie Mellon changed its password policy to something really complicated,” says Cranor, who is currently on leave from the university to serve as chief technologist at the US Federal Trade Commission in Washington DC. The university said that it was trying to conform to standard password guidelines from NIST. But when Cranor investigated, she found that these guidelines were based on educated guesses. There were no data to base them on, because no organization wanted to reveal its users’ passwords, she says. “So we said, ‘This is a research challenge.’”

### TOP 10 MOST COMMON PASSWORDS 2015

1. 123456
2. password
3. 12345678
4. qwerty
5. 12345
6. 123456789
7. football
8. 1234
9. 1234567
10. baseball

SOURCE: SPLASHDATA  
(GO.NATURE.COM/BIXMEK)



Cranor and her colleagues put a wide range of password policies to the test<sup>6</sup> by asking 470 computer users at Carnegie Mellon to generate new passwords based on different requirements for length and special symbols. Then they tested how strong the resulting passwords actually were, how much effort was required to create them, how easy they were to remember — and how annoyed at the system the participants became.

One key finding<sup>7</sup> was that organizations should forget the standard advice that complex gobbledygook words such as 0s7G0\*7j%xs\$a are safest. “It’s easier for users to deal with

## CHANGING PASSWORDS EVERY 90 DAYS RANKS BETWEEN USELESS AND COUNTERPRODUCTIVE.

password length than password complexity,” says Cranor. An example of a secure but user-friendly password might be a concatenation of four common but randomly chosen words — something like usingwoodensuccessfuloutline. At 28 characters, it is more than twice as long as the gibberish example, but much easier to remember. As long as the system guards against people making stupid choices such as passwordpassword, says Cranor, strings of words are quite hard for attackers to guess, and provide excellent security.

### TIME FOR A CHANGE

Another key finding, says Cranor, is that unless there is reason to think that the organization’s security has been compromised, the standard practice of forcing users to change their passwords on a 30-, 60- or 90-day schedule ranks somewhere between useless and counterproductive (see [go.nature.com/2vq6r4](http://go.nature.com/2vq6r4)). For one thing, she says, studies show<sup>8</sup> that most people respond to such demands by choosing a weaker password to begin with, so that they can remember it, and then making the smallest change that they can get away with. They might increase a final digit by one, for example, so that password2 becomes password3 and so on. “So if a hacker guesses your password once,” she says, “it won’t take them many tries to guess it again.”

Besides, she says, one of the first things hackers do when they break in is to install a key-logging program or some other bit of malware that allows them to steal the new password and get in whenever they want. So again, says Cranor, “changing the password doesn’t help”.

Sasse sees encouraging signs that such critiques are being heard. “For me, the milestone was last year when GCHQ changed its advice on passwords,” she says, referring to the Government Communications Headquarters, a key UK intelligence agency. GCHQ issued a public document<sup>9</sup>, containing several citations to the research literature, that gave up on long-established practices such as demanding regular password changes, and instead urged managers to

be as considerate as possible towards the people who have to live with their policies. “Users have a whole suite of passwords to manage, not just yours,” goes one bit of advice. “Only use passwords where they are really needed.”

### ATTACK THE ATTACKERS

If research can uncover weak points in user behaviour, perhaps it can also find vulnerabilities among the attackers.

In 2010, Stefan Savage, a computer scientist at the University of California, San Diego, and his team set up<sup>10</sup> a cluster of computers to act as what he calls “the most gullible consumer ever”. The machines went through reams of spam e-mails collected from several major antispam companies, and clicked on every link they could find. The researchers focused on illegal pills, counterfeit watches and handbags, and pirated software — three of the product lines most frequently advertised in spam — and bought more than 100 items. Then they used specially designed web-crawling software to track back through the spammers’ supply network. If an illicit vendor registered a domain name, made payments to a supplier or used a bank to accept credit-card payments, the researchers could see it. The study exposed, for the first time, the entire business structure of computer criminals — and revealed how surprisingly sophisticated it was.

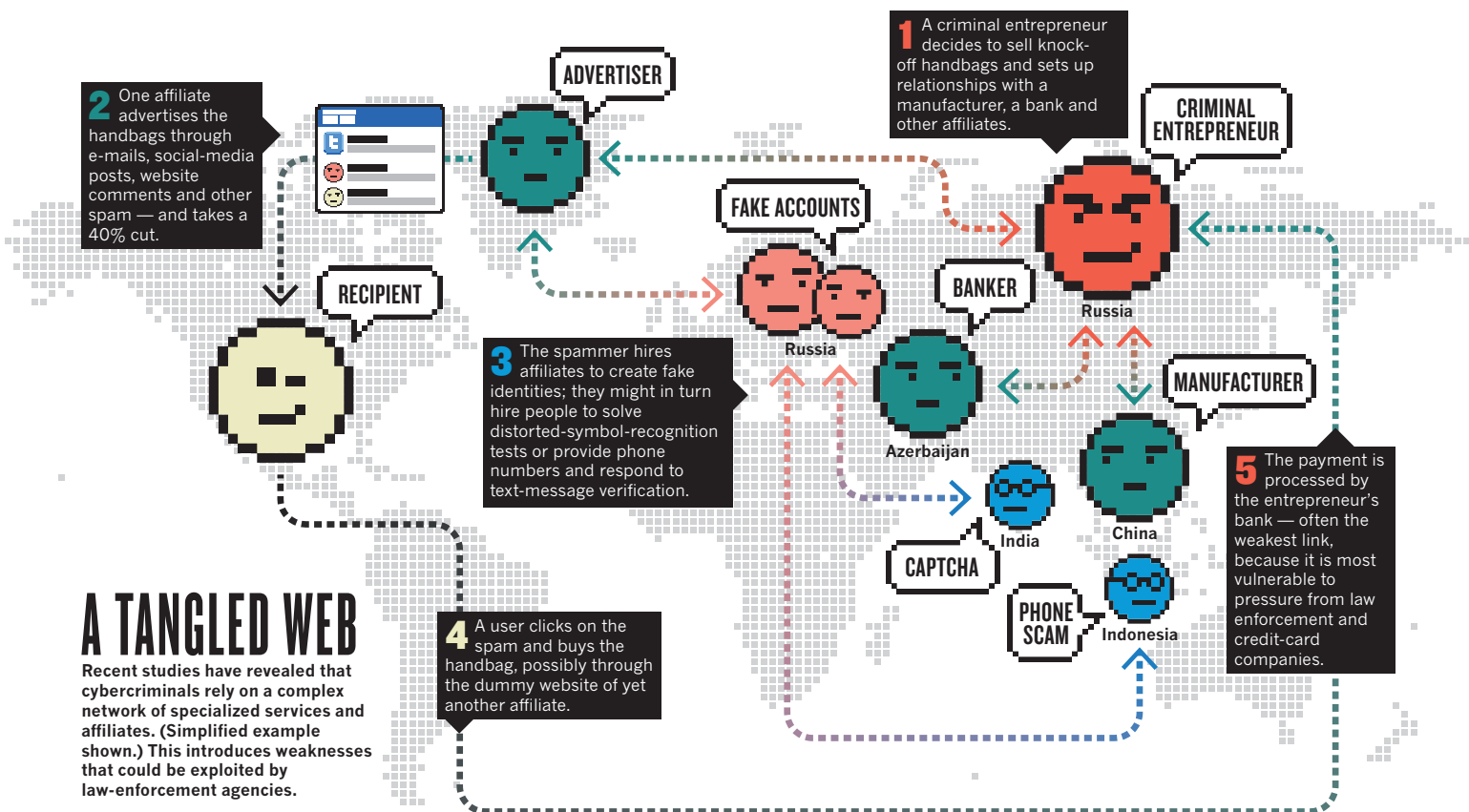
“It was the ultimate hothouse of weird new entrepreneurial ideas,” says Savage, “the purest form of small-business capitalism imaginable — because there is no regulation.” Yet there was order, even so. “Say you have a criminal activity you want to engage in,” Savage explains — for example, selling counterfeit drugs. You set up shop by creating the website and the databases, striking a deal with a bank to accept credit-card payments and creating a customer-service arm to deal with complaints — all the back-end parts of the business (see ‘A tangled web’).

“You don’t send the spam yourself,” says Savage. “You open that up to affiliates” — specialists who know how to send reams of clickable messages that fool people’s spam filters. “They get 30–40% of the purchase price for any order they bring to you,” he says. And if the brilliant idea turns out to be a dud — well, they just go and spam for someone else.

This affiliate business model has been confirmed in subsequent studies by Savage and many others<sup>11</sup>, and turns out to apply to a broad range of cybercrimes, from the sale of knock-off handbags to ransomware, credit-card piracy and other forms of cybertheft. All are supported by the same underground economy of affiliate services, of which spam generation is only one. Others range from companies in India where people spend their days typing in characters from CAPTCHA symbol-recognition tests — thus ‘proving’ that a malicious program is human — to an up-and-coming spam alternative known as search-engine poisoning, in which people who click on legitimate-looking search results are redirected to malicious websites.

Unfortunately for law-enforcement agencies, tracing the structure of this underground economy rarely helps them to arrest the individuals involved; real-world identities tend to be closely guarded behind online pseudonyms. And in any case, the criminal cyberinfrastructure is remarkably resilient. In October 2013, for example, the FBI managed to shut down Silk Road, an eBay-like website that linked buyers and sellers for illicit commodities including hard drugs. Silk Road 2.0 appeared online a month later. And when the FBI shut down that site in late 2014, still others popped up.

However, researchers have uncovered some potentially more effective ways to attack the underground economy. Savage and his colleagues found<sup>12</sup> that by far the weakest



## A TANGLED WEB

Recent studies have revealed that cybercriminals rely on a complex network of specialized services and affiliates. (Simplified example shown.) This introduces weaknesses that could be exploited by law-enforcement agencies.

SOURCE: REF. 11

links were the banks that processed credit-card payments to the profit centres. They were at the mercy of the credit-card companies, whose contracts generally state that any bank that represents a merchant must guarantee that a sale is legal — and is liable for paying customers back if they complain. Few banks were willing to take such risks. “It turned out that 95% of counterfeit spam on the planet went through just three banks,” says Savage: one each in Azerbaijan, Latvia and St Kitts and Nevis. In November 2011, Microsoft worked with Visa to pressure those banks to drop the vendors that were pirating its products. “And for 18 months,” says Savage, “there was no one selling pirated Microsoft software on the Internet.”

It was not a permanent solution, however: banking support for the shady software vendors eventually moved to east Asia, where Western companies and law-enforcement agencies have considerably less leverage. Still, the hope is that continuing research will be able to make a big difference in the long run. “For the first time,” says Nicolas Christin, a computer engineer who studies the human side of cybersecurity at Carnegie Mellon, “we have vast amounts of data about the underground economy.”

Try that in the real world, says Christin: anyone who wanted to understand, say, the street-drug trade in Pittsburgh would have to go undercover and risk getting killed. And even then, they would get only a fragmentary, ad hoc glimpse of the whole picture.

But in the online world, every transaction leaves a digital trail, says Christin, who has led research on Silk Road — especially when payments are made using digital currencies such as Bitcoin. “And for an economist, that’s wonderful.” Christin and others in this field have watched criminal systems grow, mature and be taken down — and others spring up in their place. They have watched coalitions form and dissolve, and tracked how the flow of money between criminals helps them to build trust.

“We’re just beginning to scratch the surface on the

analysis,” says Christin. But he foresees this flood of data resulting in a new fusion of computer science with social science and conventional law enforcement. “It may actually be very fruitful ground for refining and testing existing theories of criminal behaviour,” he says.

Savage has a similar hope. Whether the focus is inward- or outward-looking, he says, “There is so much snake oil around security. Very few decisions are based on data.” Continuing research could help people to base more of those decisions on evidence, he says. “But to do that, you have to look at the people involved — their motivations and incentives.” ■

**M. Mitchell Waldrop** is a features editor for *Nature* in Washington DC.

1. *Net Losses: Estimating the Global Cost of Cybercrime* (Center for Strategic and International Studies, 2014); available at [go.nature.com/15nom3](http://go.nature.com/15nom3)
2. *IBM 2015 Cyber Security Intelligence Index* (IBM, 2015); available at [go.nature.com/qcxkux](http://go.nature.com/qcxkux)
3. Bonneau, J., Herley, C., van Oorschot, P. C. & Stajano, F. *Proc. IEEE Symp. on Security and Privacy* 553–567 (2012).
4. Steves, M. et al. *Report: Authentication Diary Study* (National Institute for Standards and Technology, 2014); available at <http://dx.doi.org/10.6028/NIST.IR.7983>
5. Kirlappos, I., Parkin, S. & Sasse, M. A. *Proc. Workshop on Usable Security* <http://dx.doi.org/10.14722/usec.2014.23007> (2014).
6. Shay, R. et al. *Symp. Usable Privacy and Security (SOUPS)* (2010); available at [go.nature.com/bwuclr](http://go.nature.com/bwuclr)
7. Ur, B. et al. *login*; 51–57 (December 2012); available at [go.nature.com/koxdc3](http://go.nature.com/koxdc3)
8. Mazurek, M. L. et al. *Proc. 2013 ACM SIGSAC Conf. on Computer & Communications Security* 173–186 (2012).
9. *Password Guidance: Simplifying Your Approach* (CESG & Centre for the Protection of National Infrastructure, 2015); available at [go.nature.com/bgxre8](http://go.nature.com/bgxre8)
10. Levchenko, K. et al. *Proc. IEEE Symp. on Security and Privacy* 431–446 (2011).
11. Thomas, K. et al. *Proc. Workshop on the Economics of Information Security (WEIS)* (2015); available at [go.nature.com/4emecm](http://go.nature.com/4emecm)
12. McCoy, D., Dharmdasani, H., Kreibich, C., Voelker, G. M. & Savage, S. *Proc. ACM Conf. on Computer and Communications Security* 845–856 (2012).

## THE PASSWORD GAME

Which of these passwords is stronger?

- *iloveyou88*
- *ieatkale88*

Test your password IQ at:

➔ **NATURE.COM**  
[go.nature.com/x13ctg](http://go.nature.com/x13ctg)



# COMMENT

**NATURAL HISTORY** The labs are full but the field is empty **p.172**

**CHEMISTRY** A study of Humphry Davy's dazzling mix of personas **p.175**

**GENETICS** Siddhartha Mukherjee's history of heredity, reviewed **p.178**

**METEOROLOGY** Why did El Niño flummox long-range weather forecasters? **p.179**

ANNA TARNHUVUD



Human-embryo research is governed by a policy that aims to accommodate diverse moral concerns.

## Revisit the 14-day rule

Studies of human development *in vitro* are on a collision course with an international policy that limits embryo research to the first two weeks of development, warn **Insoo Hyun, Amy Wilkerson and Josephine Johnston**.

On 4 May, two groups reported that they had sustained human embryos *in vitro* for 12–13 days<sup>1–3</sup>. Embryos normally implant in the wall of the uterus at around day seven. Until now, no one had reported culturing human embryos *in vitro* beyond nine days<sup>4</sup>, and rarely have they been sustained for more than seven.

This latest advance comes only

21 months after the researchers at the Rockefeller University in New York City (some of whom are involved in the latest embryo-culturing work) announced that, under certain conditions, individual human embryonic stem cells can self-organize into structures akin to the developmental stages of embryos soon after implantation<sup>5,6</sup> (see “Two advances in human developmental

biology”). The cells were obtained from pre-existing stem-cell lines (derived from 4–5-day-old embryos donated through fertility clinics).

In principle, these two lines of research could lead to scientists being able to study all aspects of early human development with unprecedented precision. Yet these advances also put human developmental ►

► biology on a collision course with the '14-day rule' — a legal and regulatory line in the sand that has for decades limited *in vitro* human-embryo research to the period before the 'primitive streak' appears. This is a faint band of cells marking the beginning of an embryo's head-to-tail axis.

The 14-day rule has been effective for permitting embryo research within strict constraints — partly because it has been technologically challenging for scientists to break it. Now that the culturing of human embryos beyond 14 days seems feasible, more clarity as to how the rule applies to different types of embryo research in different jurisdictions is crucial. Moreover, in light of the evolving science and its potential benefits, it is important that regulators and concerned citizens reflect on the nature of the restriction and re-evaluate its pros and cons.

### POLICY TOOL

The 14-day limit was first proposed in 1979 by the Ethics Advisory Board of the US Department of Health, Education, and Welfare<sup>7</sup>. It was endorsed in 1984 by the Warnock committee in the United Kingdom<sup>8</sup>, and in 1994 by the US National Institutes of Health's Human Embryo Research Panel<sup>9</sup>.

In at least 12 countries, this limit is encoded in laws governing assisted reproduction and embryo research (see 'International agreement'). The rule is also embodied in numerous reports commissioned by governments, and in scientific guidelines for embryo and assisted-reproduction research. These include China's 2003 Ethical Guiding Principles on Human Embryonic Stem Cell Research and India's 2007 Guidelines for Stem Cell Research and Therapy.

Some versions of the rule cover embryos created by any means; others apply only to products of fertilization. Some explicitly refer to gastrulation (when three different cell layers appear) or the formation of the primitive streak; others mention only the 14 consecutive days of development. In most cases, however, what seems to be crucial is the stage of development that the 14th day typically represents, not the consecutive number of days in culture.

The formation of the primitive streak is significant because it represents the earliest point at which an embryo's biological individuation is assured. Before this point, embryos can split in two or fuse together. So some people reason that at this stage a morally significant individual comes into being.

Yet views differ on the moment in development at which a human embryo obtains sufficient moral status that research on it should be prohibited. Some, for instance, believe that the cut-off is the point of

fertilization; others argue that it comes much later, when the embryo develops into a fetus that can experience pain, exhibit brain activity or survive outside the womb.

Revisiting the 14-day rule might tempt people to try to rationalize or attack the philosophical coherence of the limit as an ethical tenet grounded in biological facts. This misconstrues the restriction. The 14-day rule was never intended to be a bright line denoting the onset of moral status in human embryos. Rather, it is a public-policy tool designed to carve out a space for scientific inquiry and simultaneously show respect for the diverse views on human-embryo research.

In fact, as a public-policy instrument, the 14-day rule has been tremendously successful. It has offered a clear and legally enforceable stopping point for research, because the primitive streak can be visibly identified and it is possible to count

the number of days that an embryo has been cultured in a dish. The alternatives at each extreme — banning embryo research altogether or imposing no restrictions on embryo use — would not have made for good public policy in a pluralistic society.

### TWO GOALS

Scientific advances are now prompting re-evaluations of other long-established research policies. For instance, it has proved difficult to maintain a previous consensus among funders, regulators and researchers that genetic engineering of human cells is permissible as long as those cells are not sperm, eggs or embryos. The clinical use of mitochondrial-replacement therapies — which cause heritable changes to future generations — was approved last year by the UK government, and deemed 'ethically permissible' earlier this year by a committee of the US Institute of Medicine.

## STREAKING AHEAD

### Two advances in human developmental biology

In 2014, researchers at the Rockefeller University in New York City placed human embryonic stem cells on plastic discs with patterned surfaces designed to support cell clustering, and treated the cells with a bone growth factor<sup>5</sup>. In one or two days, the cells had arranged themselves into radially symmetric patterns. These mirror — in flattened form — the organization of embryos soon after implantation in the uterine wall. From the outside in, concentric circles of cells form each of the three germ layers that give rise to all fetal tissues: the endoderm, mesoderm and ectoderm cells. These *in vitro* models even show evidence of primitive-streak-like regions.

These self-organizing structures, although embryo-like, are essentially two-dimensional. Other *in vitro* models have demonstrated some degree

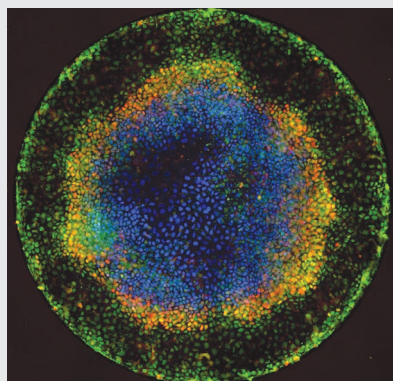
of self-organization when human embryonic stem cells are cultured in 3D environments<sup>10</sup>.

All these differ from implanted human embryos resulting from a fertilized egg in several ways. But it is plausible that researchers could one day create more comprehensive 3D models<sup>6</sup>.

This week, teams led by researchers at the Rockefeller University and the University of Cambridge, UK, report<sup>1,2</sup> that they cultured intact human embryos *in vitro* and obtained interpretable images up to day 12–13. Each team stopped its studies by day 14 in accordance with UK law and international guidelines.

The latest work provides compelling insight into how the early human embryo transitions from a floating hollow ball of cells to a three-layered gastrula attached to the uterus. Studies on human embryos sustained in culture could help to clarify whether self-organizing structures truly mimic early human development. Tracking, in real time, the morphological and molecular changes in embryonic cells and the interactions between them during these later days of development, could elucidate the cell-signalling pathways that guide embryo organization and tissue formation.

These techniques could shed light on the disorders that result in early pregnancy losses and birth defects, and facilitate clinical applications of stem-cell research. In conjunction with gene-editing tools, they could even help to determine the role of specific genes in human development.



Human embryonic stem cells form self-organized spatial patterns.

REF. 5



Some might conclude from such developments that policymakers redefine boundaries expediently when the limits become inconvenient for science. If restrictions such as the 14-day rule are viewed as moral truths, such cynicism would be warranted. But when they are understood to be tools designed to strike a balance between enabling research and maintaining public trust, it becomes clear that, as circumstances and attitudes evolve, limits can be legitimately recalibrated.

Any decision to revise the 14-day rule must depend, however, on how well any proposed changes can uphold the rule's two chief goals: supporting research and accommodating diverse moral concerns.

The rule became a standard part of embryo-research oversight through the convergence of deliberations of various national committees over decades. Hundreds of medical and scientific associations submitted recommendations, and dozens of public forums were held. Any formal changes to this rule should occur through similar processes of consensus-building involving experts, policymakers, patients and concerned citizens.

Ideally, discussion should begin at an international level given the global nature of this research — although taking local cultural and religious differences into account properly would also require national-level debates. A complication is that in many countries, a revision to the 14-day rule would involve a legislative change. Yet the kind of international discourse that we envision could facilitate and inform local decisions to amend law or research policy.

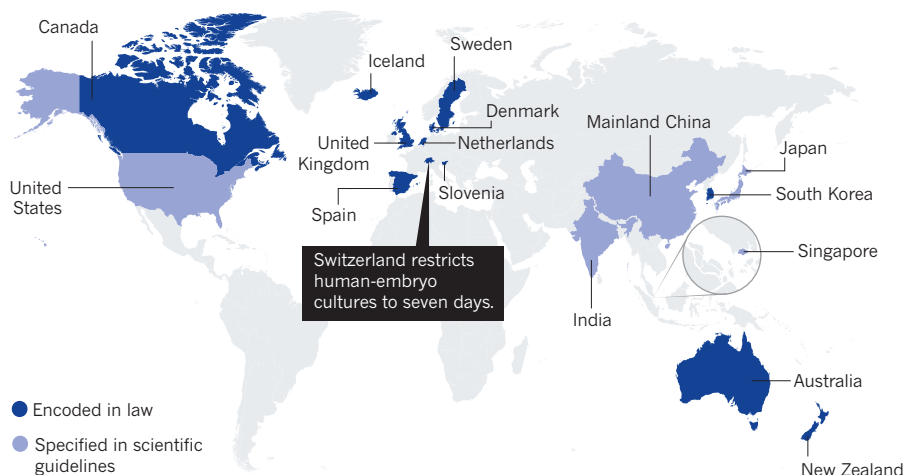
There are precedents for this type of international discourse. In response to the development of powerful gene-editing technologies such as CRISPR–Cas9, the US National Academy of Science, the US National Academy of Medicine, Britain's Royal Society and the Chinese Academy of Sciences jointly hosted an international summit in December last year to discuss scientific, ethical and governance issues raised by the research. The second component of this initiative — a science and policy review and report on human gene editing — is ongoing.

## A PATH FORWARD

Scientists have a crucial part to play in this process. In 1985, when the legality of human-embryo research in the United Kingdom was threatened by a parliamentary bill, *Nature* editors appealed to embryologists to submit explanations of their research and its importance — to educate policymakers and the public before undue restrictions on research were passed (see *Nature* 314, 11; 1985).

## INTERNATIONAL AGREEMENT

Twelve countries (dark blue) have laws that restrict *in vitro* research on human embryos to within the first 14 days of development. In five others (pale blue), nationally commissioned scientific guidelines specify the 14-day rule. Guidelines issued by the International Society for Stem Cell Research for the global research community also articulate the 14-day rule.



Today, researchers of human developmental biology should similarly engage with the public about what they are doing and why it matters. And they should consider designing their experiments in a way that, while furthering discovery, also addresses people's moral concerns.

In the immediate future, researchers should work closely with their local research-oversight committees to ensure that they are not at risk of violating current laws or guidelines. There are currently ambiguities around the legal definition of 'human embryo' in some jurisdictions, and uncertainties around the biological potential of self-organizing, embryo-like structures<sup>6</sup>.

Next week, the International Society for Stem Cell Research (ISSCR) will release its revised guidelines for stem-cell research. These guidelines are the result of a multinational, interdisciplinary task force (which included one of us, I.H.) with input from stakeholders around the world. One of the goals of these guidelines is to provide a framework for those concerned about how research oversight should proceed in light of new forms of embryo research.

In the short term, we think that the ISSCR's recommended approach to oversight of work involving human embryos offers a practical path forward — especially if supplemented with input from representatives of the many advisory committees that have adopted the 14-day rule. Obvious candidates are the UK Human Fertilisation and Embryology Authority,

the US National Academies of Sciences, Engineering, and Medicine, and the Chinese Ministry of Science and Technology and Ministry of Health.

Close collaboration between these organizations could help to prevent a public backlash and the implementation of reactive, more restrictive limits on research. ■ [SEE NEWS & VIEWS P.182](#)

**Insoo Hyun** is associate professor of bioethics and philosophy at Case Western Reserve University School of Medicine, Cleveland, Ohio, USA. **Amy Wilkerson** is associate vice-president for research support at the Rockefeller University in New York, New York City, USA. **Josephine Johnston** is director of research and a research scholar at the Hastings Center in Garrison, New York, USA.  
e-mail: [insoo.hyun@case.edu](mailto:insoo.hyun@case.edu)

1. Deglincerti, A. *et al. Nature* <http://dx.doi.org/10.1038/nature17948> (2016).
2. Shahbazi, M. N. *et al. Nature Cell Biol.* <http://dx.doi.org/10.1038/ncb3347> (2016).
3. Rossant, J. *Nature* **533**, 182–183 (2016).
4. Carver, J. *et al. Hum. Reprod.* **18**, 283–290, (2003).
5. Warmflash, A., Sorre, B., Etoc, F., Siggia, E. D. & Brivanlou, A. H. *Nature Meth.* **11**, 847–854, (2014).
6. Pera, M. F. *et al. Nature Meth.* **12**, 917–919, (2015).
7. Ethics Advisory Board, Department of Health, Education, and Welfare. *HEW Support of Research Involving Human In Vitro Fertilization and Embryo Transfer* (US Government Printing Office, 1979).
8. UK Department of Health and Social Security. *Report of the Committee of Inquiry into Human Fertilisation and Embryology* (Her Majesty's Stationary Office, 1984).
9. Ad Hoc Group of Consultants to the Advisory Committee to the Director, NIH. *Report of the Human Embryo Research Panel* (US Government Printing Office, 1994).
10. Taniguchi, N. *et al. Stem Cell Rep.* **5**, 954–962 (2015).



David Attenborough in one of his early natural-history documentaries, *Zoo Quest*, more than 60 years ago.

# Restore our sense of species

**Klaas-Douwe B. Dijkstra** has named a new dragonfly after David Attenborough to mark the broadcaster's 90th birthday — and to honour the importance of knowing the natural world.

Many of the greatest communicators have been naturalists. E. O. Wilson has his ants and Jared Diamond his birds. Oliver Sacks studied ferns as well as brains. The views of these thinkers were opened up by their exploration of the biosphere, enabling them to expose that vision to the world. One of the greatest naturalists of all is not a scholar in the traditional sense, but a broadcaster. To mark the 90th birthday on 8 May of documentary pioneer David Attenborough, I was given the honour of naming a new species of dragonfly — his favourite insect — for him (see 'David's dragonfly').

Attenborough's many hundreds of hours of exquisite television are a reminder that a primary reason to study nature is enlightenment. I am, of course, not the first to name a species for him — at least 15 animals and plants now bear his name. But doing so has led me to reflect on why we need naturalists like Attenborough today more than ever.

Although the impact of humans on all other life is beyond apocalyptic, our consciousness of its diversity is medieval. Almost 3 centuries since Carl Linnaeus came up with his naming system, perhaps only 1.2 million of an estimated 8.7 million extant eukaryote species (animals, plants, fungi and protists) are named, and more than half of them are insects such as dragonflies<sup>1</sup>. Even worse than this taxonomic deficit is our grasp of natural history. The basic ecology, distribution and status of only 80,000 species are known well enough to assess their extinction risk, and about 29% are in danger ([www.iucnredlist.org](http://www.iucnredlist.org)).

The scale of this incredible richness, ignorance and destruction is hard to fathom. Imagine for a second that each of the estimated 6.5 million terrestrial species had an equal share of the total available land on Earth. Each species' plot would cover an area only one-quarter the size of Manhattan, so the human plot could be walked around in just three hours. The 80,000 species we are familiar with would only cover an area equal to Spain, France and Turkey; and we would not even know the natural world beyond the combined areas of Europe, India and China. Yet we are set to void of life an area equivalent to the New World. The biosphere has been charted as well today as the globe was in Christopher Columbus's day, yet the biological apocalypse is already complete.

After spending 32 of my 41 years in the field, I'm still agog at life's splendour. The beauty I see as I search for dragonflies on expeditions in Gabon or watch birds around Stellenbosch on my way to work can be so absurd it makes me laugh, so diverse it makes me gasp for air, so intense that I binge-watch as if there is a

DAVID ATTENBOROUGH/BBC



cliff-hanger in every impression.

Last year, my colleagues and I described 60 new dragonfly species at once, adding 1 species to every 12 known in Africa<sup>2</sup>. Why? To show that most of what is unknown, however conspicuous, is simply not looked for. The field is empty while the labs are full. A student I know, who is passionate about exploring beetles — the most varied animal group on Earth — has ended up studying the gene expression of one model species in a lab.

Intact biodiversity provides undeniable proof that we can inhabit our environment without destroying it. Just when naturalist-taxonomists are needed most to expose the evidence, their position has become weak. We in the field are partly to blame for this marginalization, having too often emphasized the scientific and economic value of what we do, thus losing sight of its impact beyond science and the economy. Our work's greatest justification lies not in biodiversity's enormous direct contribution to human well-being, but in the moral counterweight that we can offer to life's runaway exploitation: biodiversity is the embodiment of sustainability.

### CORE VALUES

Attenborough once said that he had “never met a child who was not interested in natural history”. For most of our existence, humans were hunter-gatherers who needed to name and know other species to survive. We evolved an affinity with nature — what Wilson calls biophilia. This is probably why observing nature can be so satisfying. If nurtured, this instinct could rapidly transform society<sup>3</sup>.

The core value of natural history and taxonomy is species sense. This is a consciousness of the existence and impact of all species, from plankton to cattle and including humankind. Species make each place special and thus worth fighting for. Life is like water, a branching river system literally a genealogy: with a unique history in every separate stream, one human action can erase an irreplaceable ecosystem<sup>4</sup>. Although a hydroelectric dam might seem a sapient energy solution, it can mean ecocide to a ‘specient’ mind — one with species sense.

Whereas every human relies on this species sense, even if only by reaping the benefits of agriculture and medicine, few in society see it as their primary responsibility. With nature held hostage by our growing demands, environmental consultants and conservationists have little time left to find out who they work for.

Most worryingly, the field of biology itself has lost species sense. Biological research is an interaction between the inventory of life's diversity and the investigation of the forces shaping it. Disciplines shift

## DAVID'S DRAGONFLY

### Field and museum research reveal Madagascan beauty

Dragonflies do not help to feed us like bees and fish do; they are not feared and persecuted like mosquitoes and snakes; nor are they studied as proxies of human psyche and society like ants and apes. Their beauty and sensitivity stand for the state and needs of nature before our own. We admire dragonflies purely for what they are — the same unconditional love for nature that David Attenborough has taught us.

In few places is the creative force of nature and the destructive force of humankind more apparent than in Madagascar. Fortunately, the dragonfly *Acisoma attenboroughi*<sup>10</sup> (pictured) can be seen easily there. Kai Schütte and I

first noticed in the field and in collections that this species had been confused with its African and Asian counterparts for 174 years — a fact confirmed by DNA studies in the molecular labs at the museums of Leiden in the Netherlands and Hamburg in Germany.

The photographer Erland Nielsen joined me on a special tour for dragonfly enthusiasts earlier this year to amass images and raise funds for a booklet introducing these freshwater sentinels to the people of Madagascar for the first time. Like so many species, Madagascar's spectacular dragonflies have been ignored since the European monographs produced in bionomy's heyday in the 1950s. [K.-D.B.D.](#)



Attenborough's pintail (*Acisoma attenboroughi*).

with the advance of theory and methods (genetics to genomics is one example). Spread across all ranks of life, specializations such as entomology and botany are more stable but also more isolated and introverted. Woven together, information and theory make biology strong.

However, when competition for financial support increased half a century ago, the seemingly static soloists stood weaker. Somehow, the idea that they were old-fashioned and lacked rigour and impact became accepted, thinning the warp of expertise that bore the weft of disciplines in biology's fabric<sup>5</sup>. The strongest ‘despeciation’ of biology occurred in the 1960s to 1980s. Within a 40-year period in the United States, textbook content related to natural history decreased from two pages of every three to

just one page; related PhDs fell from two in five to one in five (even as the total number in biology tripled); and the median number of courses on natural history required for a biology bachelor's degree dropped from two to zero<sup>6</sup>.

This unravelling has also affected the ultimate custodians of species sense: natural history museums. With decreasing support, these have often increased emphasis on the separate outcomes of their taxonomic expertise without reinforcing the foundation itself. Public outreach must draw bigger crowds; scientists must chase loftier questions; and collections must focus on preservation. As a result, collections-based research has lost ground<sup>7</sup>.

The species expertise from such research may be the most impactful knowledge ►

► of biodiversity, used by enthusiasts and practitioners every day. Moreover, it connects the institutes' legacy, science and public functions. Ultimately, the dissolution of their core tasks can undermine the very survival of museums. Should they keep collections that are not used for research or outreach? Or focus on the type of research conducted at institutes that do not have collections?

Pressed by human indifference and habitat destruction, harried by bureaucracy and regulations, marginalized in education, and spurned by the science and institutes that they founded, many naturalists and taxonomists now have a deep feeling of exasperation. With less species sense, how fit are today's biologists to write research papers, undertake surveys, manage collections and teach our children?

### REVIVE BIONOMY

We need one name to acknowledge the elementary and edifying exploration of life, with its own funding mechanisms. Although venerable, 'natural history and taxonomy' can sound clunky, transient and restrictive. I like 'bionomy'. Coined 150 years ago by Ernst Haeckel as an alternative to ecology, bionomy has never been applied widely (although 'bionomics' is used occasionally to describe the natural history of species).

Like the science from which it is increasingly divorced, bionomy is a human endeavour in its own right. Arguing that biology builds on bionomy and thus owes it support is moot if science's driving force is advance and not consolidation. With independent financing, from a global pot, current indicators of impact — including journal citations and methodological innovation — can be replaced by appropriate ones such as urgency and applicability.

Most habitats will be gone before the most basic surveys are done. Mapping genomes, running models and experiments, perhaps even digging fossils, may just have to wait. Predicting future biodiversity, or studying past extinctions to understand the present, at times seem to be almost fantastical excuses for ignoring the loss itself. In an era of extinction, there are no greater priorities than to accelerate the synthesis of life, salvage knowledge and increase awareness. To do so, we need our strongest familiarity with all species.

Thanks to the technological reinvention of natural history and taxonomy, all imaginable knowledge can now be integrated and analysed<sup>8</sup>. Online repositories and genetic tools released a deluge of valuable information, although dwindling expertise struggles

to validate it. Citizen science is a powerful contributing force, because no research is closer to the public's heart, but it also needs authoritative support. We may finally learn the true extent and intricacy of biodiversity, but innovations have increased the need for what they seemed to replace by baring the enormity of the expertise gap.

Bionomers are often criticized for lacking shared and achievable goals<sup>9</sup>. Because enlightenment is even more important than information, we must invest in people before tools. Our target should not be quota of species known or access to passive data, but a volume of active experts. If we can send a probe to search for unlikely life in space, then US\$10 billion of global core funding over a decade (roughly Turkey's annual science spending) would be a bargain for humanity to develop a conscience for all life around us.

The biodiversity products needed most today are not patents or papers, but inventories, field guides, Red Lists of threatened species, teaching materials and media campaigns. Whether academics or amateur, good bionomers have initiative and drive, and only need prospects and recognition to deliver these products. Prizes that reward their typically lifelong personal investment may be the best way to double the active expertise worldwide. For instance, if a handbook took three years of funding to complete, similar funding will see more such output.

Astronomers and astronauts discover the Universe; bionomers and bionauts uncover life. Just as we feel an instant sense of our insignificance when we stare into the cosmos, or experience the exhilaration of the expanding landscape as we ascend a mountain, bionomy stretches our horizon. Each species is a world parallel to our own, invoking a sense of being among equals. That, I believe, is what Attenborough has taught us and what we must expand. ■

**Klaas-Douwe B. Dijkstra** is a dragonfly researcher affiliated with Stellenbosch University in South Africa and Naturalis Biodiversity Center in Leiden, the Netherlands.  
e-mail: african.dragonflies@gmail.com  
Twitter @bionomer

1. Mora, C., Tittensor, D. P., Adl, S., Simpson, A. G. B. & Worm, B. *PLoS Biol.* **9**, e1001127 (2011).
2. Dijkstra, K.-D. B., Mézière, N. & Kipping, J. *Odonatologica* **44**, 447–678 (2015).
3. McCord, E. L. *The Value of Species* (Yale Univ. Press, 2012).
4. Dijkstra, K.-D. B., Monaghan, M. T. & Pauls, S. U. *Ann. Rev. Entomol.* **59**, 143–163 (2014).
5. Pearson, D. L., Hamilton, A. L. & Erwin, T. L. *Bioscience* **61**, 58–63 (2011).
6. Tewksbury, J. J. *et al.* *BioScience* **64**, 300–310 (2014).
7. Winker, K. & Withrow, J. J. *Nature* **493**, 480 (2013).
8. Godfray, H. C. J. *Nature* **417**, 17–19 (2002).
9. Wheeler, Q. *Syst. Biodiv.* **8**, 11–15 (2010).
10. Mens, L. P., Schütte, K., Stokvis, F. R. & Dijkstra, K.-D. B. *Zootaxa* **4109**, 153–172 (2016).





In James Gillray's 1802 cartoon, young Humphry Davy works the bellows at a Royal Institution lecture on pneumatics.

## HISTORY

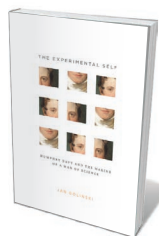
# A chemist's contradictions

Mark Peplow parses a book on Humphry Davy's dazzling mix of personas.

It is November 1799, and Humphry Davy is plastered. Fuelled by alcohol and nitrous oxide, the callow 20-year-old — who within a decade would be feted as the greatest English scientist since Isaac Newton — is seized by an epiphany. “Nothing exists but thoughts!” he cries. “The universe is composed of impressions, ideas, pleasures and pains!”

Historian Jan Golinski recounts this “unprecedented binge” in *The Experimental Self*, a study that unpicks the chemist's complexities and contradictions. It was Davy the experimentalist who prepared the ‘laughing gas’ and meticulously recorded its effects. But it was Davy the Romantic poet who gave vent to passionate pronouncements while in the grip of the gas. This amalgam of characteristics helped to establish his reputation as a swashbuckling scientist, a glamorous adventurer at the frontiers of knowledge and experience.

From relatively humble beginnings in Cornwall, Davy ascended to the presidency of the Royal Society in London, winning a knighthood and a baronetcy along the way. He discovered a slew of chemical elements, including sodium and potassium, developed a safety lamp for miners and helped to make chemistry the most fashionable science of the age. He worked at London's Royal Institution (RI), which exploited his fame to raise money, and by 1808 he had used the funding to build a monster 2,000-plate voltaic battery. It was the Large Hadron Collider of



**The Experimental Self: Humphry Davy and the Making of a Man of Science**  
JAN GOLINSKI  
University of Chicago Press: 2016.

its day, a device for carving matter into its constituent parts — and a symbol of national pride, beating the French in a race of voltaic one-upmanship.

As a gifted scientist and charismatic personality, Davy has been a magnet for biographers for two centuries. Golinski takes a new approach. *The Experimental Self* does not claim to be a comprehensive biography. Instead, it investigates the different identities that Davy constructed, along with those ascribed to him by others: enthusiast, genius, dandy, discoverer, philosopher, traveller.

Golinski's framing echoes Davy's own musings. When he was about 21, Davy wrote a piece in his notebook entitled ‘The History of Passion — A Philosophical Narrative’, detailing identities that a person might adopt throughout their life. He also wrote unpublished short stories exploring how character archetypes of



► the day — such as a student or a lover of nature — might be changed by their life experiences. Golinski cites these “experiments in selfhood” as evidence that Davy thought carefully about his character to create different personas, which he deployed strategically to build his career. He deliberately shaped his speech, manners and deportment to fit into London’s high society, and used his talents as a showman and rhetorician to attract huge audiences to his RI lectures on the latest chemical phenomena.

Golinski thus provides a fluid view of Davy, hopping back and forth between different periods of his life. It’s a refreshing approach, although it leads to some repetition when the same periods of Davy’s life (and even the same incidents) are viewed through different lenses of personality.

Davy the enthusiast developed during the nitrous oxide experiments at the Medical Pneumatic Institution in Bristol, where he drew on the Romantic aesthetic of the sublime to describe the mingled pleasure and fear that the gas triggered in himself and willing volunteers. At the time, ‘enthusiast’ meant someone who was mentally unbalanced by their passions; Davy used the word about himself to acknowledge how the gas distorted his perceptions. That work soon stirred controversy, with critics ridiculing the effects as a collective hysteria. Public shows of intoxication were lampooned by satirists such as the cartoonist James Gillray.

Once he moved to the RI in 1801, Davy took a more moderate tone. He still inhaled the gas during lectures, but did not offer it to others. Nevertheless, the lectures quickly became London’s hottest ticket. Unlike the learned Royal Society, the institution had a mission to diffuse knowledge



Humphry Davy as a young man.

to the public and exhibit applications of science. Davy’s dramatic demonstrations delivered that in spades, offering spectacle and enlightenment in equal measure. The RI’s tiered lecture theatre was soon crammed with toffs, including a remarkable number of women.

Davy the genius emerged from this period of public exposure. By emphasizing mannerisms associated with intellectual greatness — eloquence, intensity, expressive movements — he cemented his position as a leading thinker, as well as a chemical showman. Riding a wave of success, he wrote the first volume of a chemistry textbook in 1812. But *Elements of Chemical Philosophy* was a flop. By putting himself and his experiments front and centre, the book appeared to be a self-congratulatory celebration of Davy the

discoverer, and was roundly criticized by scientists such as chemist Thomas Thomson for lacking the neutrality expected of a sober overview of the field.

Increasingly, Davy’s carefully constructed personas caused problems. He became president of the Royal Society in 1820, but his charisma cut less ice there. His suggestion that women should be allowed to attend evening meetings at the society, as at the RI, was rejected, and even led some to question his masculinity. Meanwhile, a younger generation of scientists resented his dependence on public display and aristocratic patronage. They wanted to be elected to the society on merit, funded by governments rather than lords. As the age of the professional scientist dawned, Davy started to look like an anachronism.

Golinski ends with Davy the traveller. The scientist made a series of continental journeys, conducting research on anything from the chemistry of volcanoes to fresco pigments as he toured the cities of Europe. This was the full Davy roadshow, a heady blend of the scientific and Romantic: he would make poetic notes about the scenery and the weather, and faithfully record their effects on his health. Davy had a series of strokes during these trips, and died in 1829; his final book, *Consolations in Travel*, was published the next year. It features dialogues between fictional characters — including two reflecting aspects of himself — that explore philosophical questions such as the nature of the soul. “*Consolations* could be seen as his last virtuoso performance,” writes Golinski, “the last of his experiments in selfhood.” ■

**Mark Peplow** is a science journalist based in Cambridge, UK.  
e-mail: [peplowscience@gmail.com](mailto:peplowscience@gmail.com)

## BIODIVERSITY

# England’s green and well-known land

**Stuart Pimm** extols Richard Fortey’s scientific and historical portrait of a beechwood.

**F**ly into London’s Heathrow airport, and off to the northwest you will spot a sprinkle of dark-green patches along the undulating fields and hedgerows. The Chilterns are hills sitting on a chalk escarpment, and retain some woodland because of their soils, topography and underlying geology, and the special management that these demand. In 2011, palaeontologist and natural historian Richard Fortey bought 2 hectares of beechwood and bluebells here and began

a diary — a “biography”, as he puts it.

In *Wood for the Trees*, Fortey’s story unfolds over two interlaced time scales: one a calendar year, the other two millennia of recorded history. His year starts in April, witness to one of those intense English springs that can follow a long, damp winter. Like Robert Browning’s famous paean to the season (1845’s

**NATURE.COM**  
For more on science  
in culture see:  
[nature.com/  
booksandarts](http://nature.com/booksandarts)

‘Home-Thoughts, from Abroad’), which began, “Oh, to be in England/Now that April’s there”, Fortey’s text is stiff with the names of trees, flowers and birds. There’s a pattern here: a very British predilection for natural history. This means that British flora and fauna are exceptionally well documented. I took this for granted as a young naturalist in Derbyshire, only to get a rude shock when looking for field guides in other countries. A well-illustrated guide to the native trees of South Florida (where



ROB FRANCIS

I write this) was not published until 2014.

It's easier to know the names of species in island ecosystems, of course. They have fewer species than those of larger land-masses; many continental flora and fauna are excluded by what William Shakespeare called Britain's "moat defensive". There are (arguably) around 30 native British trees, for example. In any case, the passion for naming everything seems uniquely British.

As the year unfolds, Fortey names familiar birds, trees and flowers, butterflies — speckled wood (*Pararge aegeria*) and peacock (*Aglais io*) — dormice and squirrels. He also identifies moths, land snails, spiders, slime moulds, flies, beetles, mosses, fungi, lichens — and three species of pseudoscorpion. Globally, taxonomists have described only small fractions of the species in these taxa. Fortey's wood is home to more than 300 species of fungi, including the familiar stinkhorn fungus, graced with the unforgettable scientific name of *Phallus impudicus*.

He notes his sources for identification. For those of us who still struggle with tools such as dichotomous keys, there are picture books, which we can scan for an image of the unknown species. William Keble Martin's 1965 *The Concise British Flora in Colour* (Ebury) is the landmark. The product of decades of painstaking illustration, ignored by presses worried about the cost of colour reproduction, it became an instant best-seller. In one photo, we see Fortey perusing illustrations to help him to identify a moth at a light trap. He asks colleagues from London's Natural History Museum, where he worked as a palaeontologist, to visit. He brings in experts on flies and canopy insects, and others to listen to the ultrasound of resident bats.

The history of Fortey's beechwood begins in Roman times and runs through the Anglo-Saxon, Viking and Norman invasions, providing an exceptionally detailed record. This is an intensively managed landscape for which the names and personal histories of centuries' worth of owners are known. The wood survives because it was a working wood. English yew trees provided the bows for archers at the Battle of Agincourt in 1415. Ash made tool handles; cherry, furniture. The woodworkers had names, too: bodgers, turners and the sawpit operators' top dogs and underdogs.

A yet longer history is the geological record. Over the 200 kilometres from

**"The wood survives because it was a working wood. Ash made tool handles; cherry, furniture."**



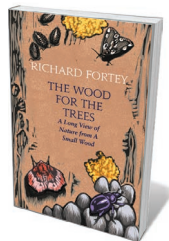
The view across the Assendon Valley to Richard Fortey's piece of woodland in England's Chiltern Hills.

Derbyshire to London lie 250 million years of rocks, from the lower Carboniferous period, 300 million years ago, to the Eocene epoch, around 50 million years ago. They sit gently on top of each other, in a perfect sequence, with hills such as the Chilterns marking out major transitions. Without this elegant simplicity and a deep understanding of the natural history that it shapes, would William Smith (see *Nature* 520, 294; 2015) have produced the first geological map?

An intermediate history is how climate is disrupting

nature. From the Victorian era onwards, legions of amateur enthusiasts have done work on species distribution that provides vital benchmarks. This ability to name so many species has meant that Britain has produced the world's largest share of studies of how plants flower earlier now than in the past and how butterflies live further north — feeding on different plants as they change habitats. And that is the true value of all these lists. As Fortey puts it: "Think of it as not so much an inventory as a catalogue leading to compelling and interlocking stories." ■

**Stuart Pimm** is professor of conservation at the Nicholas School of the Environment at Duke University in Durham, North Carolina, and directs the non-profit organization SavingSpecies. e-mail: [stuartpimm@me.com](mailto:stuartpimm@me.com)



**The Wood for the Trees: The Long View of Nature from a Small Wood**  
RICHARD FORTHEY  
William Collins: 2016.



# On the heredity trail

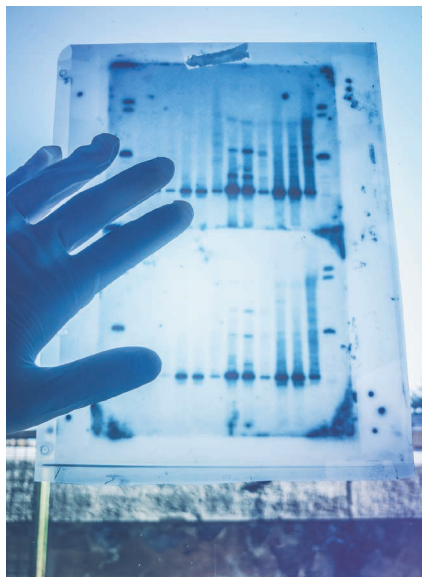
Matthew Cobb assesses Siddhartha Mukherjee's history of human genetics.

In 2011, Siddhartha Mukherjee won a Pulitzer prize for *The Emperor of All Maladies* (Scribner, 2010), which intertwined science and his own experience as an oncologist. In *The Gene*, Mukherjee uses a personal approach to describe our understanding of heredity. Despite its subtitle ('An Intimate History'), the historical sections of *The Gene*, ranging from 1860 to the present, are not intended to show the convoluted route to current knowledge. They are primarily a tool for explaining the basics of medical genetics.

As a consequence, the complexities of the past are ironed out. Discovery is presented not as a messy reality full of dead ends, but as a linear thread leading inexorably to today. Conclusions of past experiments are presented in terms of modern understanding, rather than as a way to explore confused contemporaneous interpretations. This is a road often followed by scientists and clinicians who write history; it irritates historians, who know that the past was more complicated.

The first half of the book takes us up to the late 1960s and presents familiar, sometimes erroneous versions of past milestones. For example, nineteenth-century genetics pioneer Gregor Mendel appears as a lone genius. In fact, his work was part of a long-term interest in heredity on the part of Cyrill Napp, abbot of the monastery in Brno — now in the Czech Republic — where Mendel was a monk. This interest was prompted by the desire of local agriculturalists (including some at the monastery) to improve their animals and plants, and began nearly 20 years before Mendel planted his peas. Sometimes, the drama is downplayed. A brief footnote describes Vernon Ingram and Francis Crick's ground-breaking 1950s demonstration that the difference between normal and sickle-cell haemoglobin is caused by a single-base difference in the relevant gene. And one of the most exciting moments in genetics history — the 1960s discovery of how genes encode proteins — is passed over in a couple of lines.

The writing comes alive in the book's second half, covering the 1970s onwards, and introduced by Enlightenment poet Alexander Pope's line: "The proper study of mankind is man". Mukherjee, as a physician, rightly takes that declaration as his own. Here, the book does become intimate. Mukherjee's account of the development of biotechnology companies in the 1970s is enriched by personal recollections from Nobel-prizewinning biochemist Paul Berg, in whose laboratory



DNA analyses promise to change health care.

Mukherjee worked in the 1990s. The passages that describe patients with genetic diseases are full of the compassion that we would all wish from our doctors. At other points, Mukherjee brings in examples from his own family, in particular his uncle and cousin, who both had schizophrenia, to frame the narrative and form the starting point for his examination of the role of genetic factors in disease.

Mukherjee's account of how the genetic basis of Huntington's disease was discovered is particularly effective, covering the personal tragedies and the motivation of the scientists who made this breakthrough, including Nancy Wexler, whose mother died of the disease. There is an atmospheric description of Wexler's fieldwork region in Venezuela, where almost 10% of the population has Huntington's disease. An endnote candidly admits that these passages were inspired not by a visit to the region as I had imagined, but by a BBC *Newsnight* report about Wexler that can be seen on YouTube. Some writers are not so honest about pulling aside the curtain of creation.

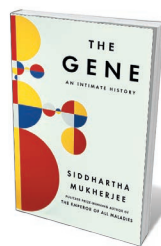
However, one consequence of Mukherjee's medical focus is that some of the most stunning discoveries

from human genomics barely get a mention. *Homo sapiens'* interbreeding with Neanderthals is allotted two sentences, and our species' sexual encounters with the mysterious Denisovans are not mentioned. The genetic legacy of those interactions may help to explain aspects of epidemiology: there seem to be links between Neanderthal DNA and a number of immunological, dermatological and psychiatric conditions, such as depression and the skin lesions called actinic keratosis. Furthermore, because the book centres on medical genetics, anyone expecting an exploration of the state of genetics as a whole will be disappointed. *Our Genes* would have been a more appropriate title than *The Gene*.

A final section examines what Berg described to Mukherjee as "the future of the future" — the amazing possibilities for manipulating the human genome that are within our grasp. Mukherjee outlines the rise and fall of gene therapy in the 1990s, always with a clinician's compassion for the tragic stories behind the technology; and discusses the potential for gene modification with tools such as CRISPR-Cas9. This section concludes with some of what Mukherjee does best, combining stories of real patients with the ethical dilemmas raised by their conditions — in this case, what would happen if their disorders were the subject of prenatal or pre-implantation testing?

*The Gene* finishes with a manifesto for living in a post-genomic world. Mukherjee provides a 13-point 'opening salvo' that outlines our knowledge of human genetics and its implications. He concludes blandly: "we need new biological, cultural, and social precepts to determine which genetic interventions may be permitted or constrained, and the circumstances in which these interventions become safe or permissible". More-consistent grappling with the ethical, philosophical and historical debates that have swirled around these issues for decades might have led to a more detailed hint of what those precepts might be. This vagueness is frustrating: Mukherjee's visceral and thought-provoking descriptions of the horrors of early-twentieth-century US eugenics clearly show what he is capable of in this regard, both as a writer and as a thinker. ■

Matthew Cobb is professor of zoology at the University of Manchester, UK. His latest book is *Life's Greatest Secret*. e-mail: cobb@manchester.ac.uk



**The Gene: An Intimate History**  
SIDDHARTHA  
MUKHERJEE  
Scribner: 2016.



# Correspondence

## Reminder to deposit DNA sequences

As members of the Advisory Committee to the International Nucleotide Sequence Database Collaboration (INSDC), which includes the DNA Data Bank of Japan (DDBJ), European Nucleotide Archive (ENA) and GenBank databases, we wish to remind the research community of the importance of depositing complete DNA-sequence data in these databases on publication of their results (see also S. L. Salzberg *et al.* *Science* <http://dx.doi.org/10.1126/science.aaf7672>; 2016). Indeed, most journals demand a database accession number as a condition of publication.

Access to the INSDC's databases is free and unrestricted (G. Cochrane *et al.* *Nucleic Acids Res.* **44** (D1), D48–D50; 2016), enabling researchers to plan experiments and to analyse existing data. As original contributions, deposited data form part of the scientific record and are citable in the literature. Authors can also correct and update their data: these amended records may be removed from the next database release, but still remain permanently available by accession number.

The INSDC has also created major new repositories for large data collections, notably the Sequence Read Archive at the National Center for Biotechnology Information (NCBI), the DDBJ Sequence Read Archive and the ENA at the European Molecular Biology Laboratory's European Bioinformatics Institute (EMBL-EBI). These repositories archive raw data from sequencing experiments, a crucial facility for reproducibility and reuse.

For papers dependent on sequence data from human subjects, unrestricted data release may not be possible. In these cases, we would encourage journal editors to insist on data sharing through other repositories that are not

part of the INSDC, such as the NCBI's Database of Genotypes and Phenotypes, EMBL-EBI's European Genome-phenome Archive or DDBJ's Japanese Genotype-phenotype Archive. **Steven L. Salzberg\*** *Johns Hopkins University, Baltimore, USA.* [salzberg@jhu.edu](mailto:salzberg@jhu.edu)

*\*On behalf of 9 correspondents (see [go.nature.com/tvllbd](http://go.nature.com/tvllbd) for full list).*

## El Niño dons winter disguise as La Niña

Seasonal weather forecasting relies almost exclusively on El Niño, the climate phenomenon associated with warming in the equatorial Pacific Ocean. When the strongest El Niño on record developed last autumn, it offered an opportunity to showcase decades of investment and advances in long-range forecasting (see *Nature* **529**, 267–268; 2016). Surprisingly, however, actual winter weather events were the opposite of those predicted.

For example, southern California's winter was more about heatwaves and wildfires than deluges; Seattle in Washington endured the wettest winter on record rather than a worsening drought; and the upper Mississippi valley experienced flooding to an extent that had previously occurred only in summer.

El Niño should strengthen the side of the jet stream that is nearer to the Equator, bringing wet weather to the southwest United States and cool temperatures to the southeast. Instead, the side nearest to the pole strengthened. This brought weather that would be more expected of La Niña, the opposite phase of the Southern Oscillation that results from cooling waters.

I suggest considering this inaccurate El Niño forecast in the wider context of the Arctic's influence. Low Arctic sea ice and high Eurasian snow cover this autumn increased a Siberian high-pressure system and heat

transport towards the North Pole, weakening the polar vortex this winter (see J. Cohen *et al.* *Nature Geosci.* **7**, 627–637; 2014). The atmospheric responses to Arctic 'amplification' were better predicted than were those to El Niño outside the tropics (see [www.aer.com/winter2016](http://www.aer.com/winter2016)).

**Judah Cohen** *Atmospheric and Environmental Research, Lexington, Massachusetts, USA.* [jcohen@aer.com](mailto:jcohen@aer.com)

## Reform oversight of Italy's science funds

We wish to highlight the stark contrast between the hardship in Italy's publicly funded research community (see G. Parisi *Nature* **530**, 33; 2016) and the reportedly largely unused sums of money allocated to a single research institute over the past 12 years.

After 3 years without any financial provision for bottom-up research, Italy's government is providing €31 million (US\$36 million) a year for the next 3 years to cover research in the humanities as well as science. Of this year's 4,431 grant applications for this modest sum, just 300–500 will be successful.

By contrast, the government plans to inject €150 million a year for the next 10 years into the Human Technopole project, which will focus on genomics, big data, ageing and nutrition. The recipient is the Italian Institute of Technology in Genoa, which is self-governing and so not publicly accountable — despite the large sums of public money involved. A recent report indicates that half of the institute's funds remained unspent in 2010–14 (see [go.nature.com/vouywf](http://go.nature.com/vouywf); in Italian).

The government should establish an adequately funded agency that has transparent jurisdiction over the funding and execution of research. The agency would also monitor the progress of the Human Technopole and oversee its accountability. (See also [go.nature.com/hgfpj](http://go.nature.com/hgfpj).)

**Ernesto Carafoli\*** *Venetian Institute of Molecular Medicine, Padua, Italy.*

**Cesare Montecucco\*** *University of Padua, Italy.*

[ernestocarafoli@gmail.com](mailto:ernestocarafoli@gmail.com)

*\*Supported by 13 signatories (see [go.nature.com/3xxovu](http://go.nature.com/3xxovu) for full list).*

## Regulate devices for brain stimulation

We are concerned that the rapid development and increased accessibility of non-invasive technologies with alleged brain-enhancing capabilities is allowing commercial interest to outpace regulatory mechanisms (see *Nature* **531**, 283–284; 2016 and *Nature* **531**, S6–S8; 2016).

Only limited technical ability is required to build a brain-stimulation device at home — or to dress it up and market it commercially. Even though electrical current can endanger cardiovascular and neural function (see [go.nature.com/ej3kgx](http://go.nature.com/ej3kgx)), there are currently no requirements for safety or efficacy testing of home-use devices through clinical-style trials. Beyond the safety of the devices themselves, the impact of regular or sustained personal use of brain stimulators is unknown.

The public may not appreciate that companies are subject to a strict regulatory framework if their product claims to help an individual to achieve normal function (that is, a treatment), but not if it is sold to enhance function. We urge governments to align their regulatory standards for both applications.

**Olivia Carter, Jason Forte** *University of Melbourne, Australia.* [ocarter@unimelb.edu.au](mailto:ocarter@unimelb.edu.au)

### CONTRIBUTIONS

Correspondence may be sent to [correspondence@nature.com](mailto:correspondence@nature.com) after consulting the guidelines at <http://go.nature.com/cmchno>.

## HUMAN EMBRYOLOGY

# Implantation barrier overcome

The early stages of human development are normally hidden within the womb, but improved techniques for culturing embryos from the blastocyst stage promise to make these steps easier to investigate. [SEE LETTER P.251](#)

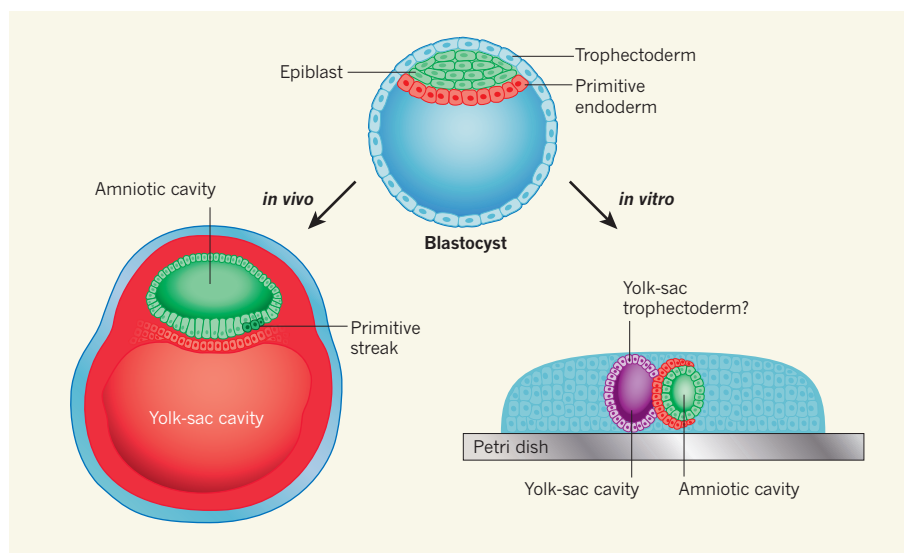
JANET ROSSANT

Studying early development in human embryos is a challenge. Few embryos are available, and research is subject to considerable ethical and legal constraints. But understanding early-stage development is vital for improving reproductive technologies, enhancing stem-cell cultures for regenerative medicine and examining early pregnancy losses. In two papers, Deglincerti *et al.*<sup>1</sup> (page 251 of this issue) and Shahbazi *et al.*<sup>2</sup> (in *Nature Cell Biology*) report that human embryos derived from *in vitro* fertilization (IVF) can self-organize in a Petri dish, forming the founding cell lineages of the fetus and its supporting tissues. This is a first step towards a clearer view of the beginnings of human life.

In mammals, including humans, a fertilized egg undergoes a series of cell divisions over the first days of development, leading to the formation of a structure called the blastocyst (Fig. 1). The first cell-lineage decisions are made at this stage, with a lineage called the epiblast, which goes on to form the entire fetus, becoming separated from two lineages that will produce non-embryonic tissues — first the trophectoderm and then the primitive endoderm. The trophectoderm gives rise to cells that form most of the placenta, whereas the primitive endoderm forms some layers of the yolk sac, which is required for early fetal blood supply.

The mechanisms that underlie blastocyst-stage lineage specification are well understood in mice, and it had been assumed that these pathways are evolutionarily conserved. However, that assumption has been challenged<sup>3</sup>. Although many of the genes that direct lineage decisions in mouse embryos are expressed in the same lineages in humans, the timing of onset and the upstream pathways that regulate their expression differ between the species<sup>4,5</sup>.

The blastocyst becomes implanted in the lining of the uterus just five days after fertilization in mice and seven days after in humans. This is a vital period, in which trophectoderm-derived cells begin to interact with the uterus, and the embryo progresses towards perhaps the most crucial step in development — gastrulation, in which an epiblast-derived cell mass called the primitive streak gives rise to the three basic cell layers from



**Figure 1 | Human embryo growth *in vivo* and *in vitro*.** During early human embryonic development, cells form a structure called the blastocyst, which is comprised of three lineages — the epiblast, which will form the fetus, and the trophectoderm and primitive endoderm, which support embryonic growth. *In vivo*, by around 12 days post-fertilization, the blastocyst has implanted in the uterus and undergone the first cell-lineage decisions. The epiblast forms an amniotic cavity and a cell mass called the primitive streak, which will produce the body's three major tissue layers. Cells derived from the primitive endoderm form a yolk sac, which is involved in early blood supply. Trophectoderm-derived cells form external structures. Deglincerti *et al.*<sup>1</sup> and Shahbazi *et al.*<sup>2</sup> cultured human blastocysts *in vitro*. Similar structures and cavities form, although their spatial relationships differ from those *in vivo*. In addition, Deglincerti and colleagues observed a previously unidentified cell type, which they dub yolk-sac trophectoderm, although its origin is unclear. (Cultured embryo adapted from ref. 1.)

which every bodily structure is derived.

In mice, signals from the primitive endoderm and trophectoderm initiate formation of the primitive streak<sup>6</sup>. But in humans, this period of development has been a complete black box. The only available information has come from rare cross-sections cut through human embryos and from non-human primates, such as rhesus monkeys. Those studies<sup>7,8</sup> show that there are major differences between primate and mouse development as the embryo implants in the uterus. Most notably, the mouse epiblast forms a cup-like structure, on one side of which form the primitive streak and amniotic folds (which will later form the fluid-filled amniotic membranes). By contrast, the primate epiblast first forms a central amniotic cavity and then flattens out to form a disc, from which the primitive streak arises at one end (Fig. 1). The spatial

relationships between lineages therefore differ between species.

The development of a strategy for culturing human embryos *in vitro* over the early post-implantation period could improve our understanding of the significance of these differences. Using a system developed for culturing mouse embryos<sup>9</sup>, Deglincerti *et al.* and Shahbazi *et al.* did just that, culturing human embryos derived from IVF up to a stage equivalent to 13 days post-fertilization *in vivo*. An improved culture medium and a better substrate for embryo attachment seem to have been the key to these advances.

The groups report that blastocysts attach to the dish, the trophectoderm spreads out and shows signs of differentiation into specialized placental cell types, and the primitive endoderm segregates from the epiblast. Shahbazi and colleagues found that a small central



cavity develops within the epiblast, apparently because the lineage reorganizes into a radially polarized structure. This is reminiscent of how the amniotic cavity is thought to form in both human and rhesus-monkey embryos<sup>7,8</sup>, although the absence of good molecular markers of amniotic tissue precludes a conclusive identification of this structure.

Both groups also observed a second cavity in the spreading primitive endoderm, which they equate to the yolk-sac cavity. Deglincerti *et al.* report that the cells lining this cavity express genes characteristic of trophoblast-derived cells. The authors suggest that this is a previously unidentified cell type, which they name yolk-sac trophoblast. However, comparison with anatomical descriptions<sup>7</sup> suggests that these cells are more likely to be derived from the primitive endoderm — perhaps with a gene-expression profile that differs from that in mice. Extension of the cultures beyond 12 days led to cavity collapse and disorganized development, although trophoblast differentiation continued.

Although these studies represent steps towards a firmer understanding of human development over the implantation period, there are many limitations still to overcome. The cultured embryos are largely flattened and two-dimensional, and so are clearly imperfect models of normal 3D embryonic development. In addition, unequivocally identifying cell types, cavities and structures in the cultures is challenging. Genome-wide expression analysis of these features might help to refine the system.

This culture method could enable researchers to probe the role of signalling molecules from the extraembryonic tissues in patterning the epiblast. By comparing these results with the signalling molecules detected in embryonic-stem-cell cultures that mimic events of gastrulation<sup>10</sup>, we might better understand how to induce human stem cells to differentiate into cell types that have therapeutic potential. The development of a 3D blastocyst culture system, akin to the 'organoid' systems used to model more-mature tissues, could also be informative. If the topological relationships between the different cell types were more normal in such 3D cultures, this might enable gastrulation to occur *in vitro*.

Currently, human embryo cultures are restricted, by international agreement, to 14 days of growth or the beginning of primitive-streak formation, whichever comes first. If gastrulation were achievable *in vitro*, what would be the impact on this 14-day rule? Improved and longer cultures could provide important information for basic human biology, improving IVF success rates and the understanding of stem-cell differentiation. However, the development of such culture systems would again raise the question of where to place the ethical limits on human embryo development *in vitro*. ■ [SEE ALSO COMMENT P.169](#)

**Janet Rossant** is in the Peter Gilgan Centre for Research and Learning, The Hospital for Sick Children, Toronto, Ontario M5G 0A4, Canada, and in the Department of Molecular Genetics, University of Toronto.  
e-mail: janet.rossant@sickkids.ca

1. Deglincerti, A. *et al.* *Nature* **533**, 251–254 (2016).
2. Shahbazi, M. N. *et al.* *Nature Cell Biol.* <http://dx.doi.org/10.1038/ncb3347> (2016).
3. Rossant, J. *Development* **142**, 9–12 (2015).
4. Blakeley, P. *et al.* *Development* **142**, 3151–3165 (2015).

5. Petropoulos, S. *et al.* *Cell* <http://dx.doi.org/10.1016/j.cell.2016.03.023> (2016).
6. Tam, P. P., Loebe, D. A. F. & Tanaka, S. S. *Curr. Opin. Genet. Dev.* **16**, 419–425 (2006).
7. Enders, A. C., Schlafke, S. & Hendrickx, A. G. *Am. J. Anat.* **177**, 161–185 (1986).
8. Luckett, W. P. *Am. J. Anat.* **144**, 149–167 (1975).
9. Bedzhov, I., Leung, C. Y., Bialecka, M. & Zernicka-Goetz, M. *Nature Protocols* **9**, 2732–2739 (2014).
10. Warmflash, A., Sorre, B., Etoc, F., Siggia, E. D. & Brivanlou, A. H. *Nature Meth.* **11**, 847–854 (2014).

This article was published online on 4 May 2016.

## ORGANIC CHEMISTRY

# Precision pruning of molecules

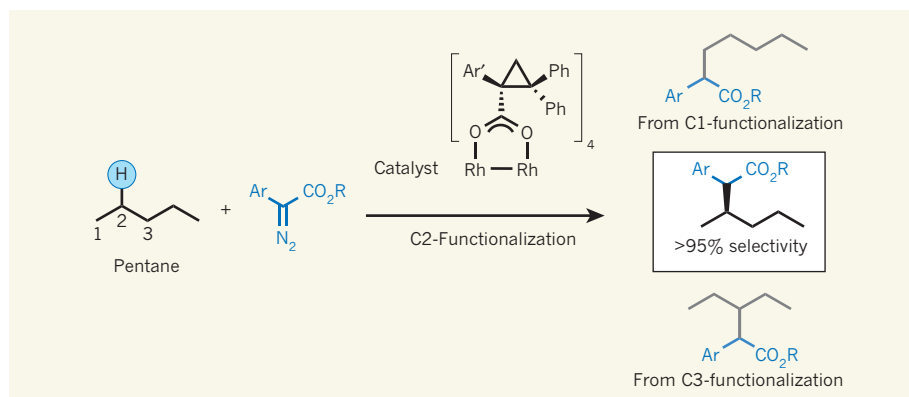
If organic molecules were trees, then the numerous carbon–hydrogen bonds within them would be leaves. A catalyst that targets one 'leaf' out of many similar other ones looks set to be a huge leap for synthetic chemistry. [SEE LETTER P.230](#)

KIN S. YANG & KEARY M. ENGLE

For much of the past century, organic chemists have focused on the reactions of functional groups: arrangements of atoms that have a characteristic, predictable reactivity profile. Such reactions have proved to be powerful tools for assembling structurally complex molecules for use in applications ranging from human medicines to pesticides, but they sometimes introduce inefficiencies that increase waste and costs. To expedite synthetic endeavours, attention has thus turned to a new generation of reactions — known as C–H functionalization reactions

— that convert previously neglected carbon–hydrogen (C–H) bonds into useful reactive functional groups. A key question is how to precisely target a single C–H bond for reaction among so many others. On page 230 of this issue, Liao *et al.*<sup>1</sup> report an outstanding advance that addresses this problem.

Carbon–carbon (C–C) and C–H bonds are the most common connecting units in organic molecules; if a molecule is viewed as a tree, then C–C bonds link together to make up the trunk and branches, which C–H bonds adorn as leaves. C–H bonds are so widespread that organic chemists almost always omit them when drawing chemical structures, instead



**Figure 1 | A selective C–H functionalization reaction.** Liao *et al.*<sup>1</sup> report a catalyst that can differentiate between similar carbon–hydrogen (C–H) bonds in hydrocarbons, such as pentane, so that the C–H bond at the second carbon atom reacts selectively with a diazo compound (blue) — a C2 functionalization reaction. This occurs with more than 95% selectivity, even though many other similarly reactive C–H bonds (not shown) are present in the molecule. Small amounts of products from C1 and C3 functionalization are also formed. Ar is a 4-substituted phenyl group or a 6-chloropyridin-3-yl group (see Fig. 4 of the paper<sup>1</sup> for molecular structures); R is  $\text{CH}_3$  or  $\text{CX}_2\text{CH}_2$ , where X is fluorine, bromine or chlorine; Ar' is 3,5-di(4-*t*-butylphenyl)phenyl; Rh, rhodium; Ph, phenyl.

showing only the skeleton of carbon atoms and heteroatoms (those that are neither carbon nor hydrogen). The fact that C–H bonds are not typically drawn also speaks to another, more subtle truth: they do not usually react in classic organic transformations.

This lack of reactivity is because the bonds are strong, are non-acidic and do not have inherent affinities for other chemical entities<sup>2,3</sup>. Collectively, these issues pose a difficulty: C–H bonds will seemingly be cleaved only by energetic reagents or under 'harsh' conditions (such as high temperatures), but such reactions are unlikely to be selective for a single C–H bond. For example, the combustion of hydrocarbons, a powerful process used to produce heat since the rise of human civilization, non-selectively cleaves all C–H and C–C bonds in the starting material to produce carbon dioxide and water. Reactions that achieve controlled and selective functionalization of C–H bonds are akin to carefully pruning a single leaf from a tree and grafting on a new branch at that position. Such transformations are underdeveloped, yet hold the promise of permanently changing how chemists design and synthesize molecules.

More than 50 years ago, inorganic and organometallic chemists discovered that transition metals interact with C–H bonds in unique ways, enabling ready C–H functionalization<sup>2,3</sup>. More recently, organic chemists have become interested in this reactivity with an eye to developing selective reactions of C–H bonds<sup>4,5</sup>. Existing strategies for obtaining such selective reactions generally involve an approach called substrate control, which depends on the intrinsic structural features and reactivity patterns of molecules. It remains a tremendous challenge to develop catalyst-controlled reactions that distinguish the subtle steric and electronic differences between C–H bonds in molecules that lack any functional groups (steric effects are those associated with the spatial crowding of chemical groups and atoms). Liao *et al.* tackled this issue directly by investigating the selective functionalization of pentane, a simple molecule that contains only C–C and C–H bonds.

Pentane contains a chain of five carbon atoms and, because of molecular symmetry, it has three distinct positions at which reactions can occur: the chain ends; the carbon atoms next to the ends (also known as the C2 positions); and the middle carbon atom. The authors sought to design a catalyst that would promote a highly selective reaction at a C–H bond at C2, a formidable undertaking (Fig. 1).

The same research group had previously pioneered a strategy for taming the reactivity of dirhodium carbenoids, a class of organometallic complex, by attaching electron-donor and electron-acceptor substituents to them<sup>6,7</sup>. The resulting species are still highly reactive, but are long-lived enough to participate in reactions with other molecules and to selectively target activated C–H bonds (such as those next to a C–C double bond, a benzene

ring or an oxygen atom). Because pentane contains only unactivated C–H bonds, a strategy is needed to enable these catalysts to distinguish between the slight differences in steric and electronic properties associated with the possible reaction sites. This means that the size, shape and electronic environment around the reactive metal centres in the complexes need to be finely tuned.

The authors therefore tested a number of catalysts for reactivity and selectivity in the C–H activation of pentane. Synthesizing a library of new catalysts for screening is often a bottleneck in reaction optimization. Catalysts tend to consist of organic ligand molecules bound to metals, and so, for each catalyst, chemists generally synthesize the ligand first and add the metal in a subsequent step.

To expedite this typically tedious process, Liao *et al.* used an ingenious approach. They first synthesized a versatile dirhodium precursor complex, from which new catalysts could be prepared at a single stroke — streamlining catalyst discovery in a process that parallels methods used for discovering small-molecule drugs. In this way, the authors discovered a catalyst that preferentially functionalizes a C–H bond at the second carbon of pentane rather than C–H bonds at the first or third carbons, forming a new C–C bond at that carbon with more than 95% selectivity. This is a truly remarkable accomplishment, given the similarity between the three C–H bonds.

The authors went on to show that their reaction is similarly effective with other saturated hydrocarbons, and with some simple compounds that contain potentially reactive functional groups such as halides, silanes and

esters. Moreover, the transformation is highly enantioselective (it yields only one of the two possible mirror-image isomers of the product). Enantioselective reactions are crucial for the synthesis of many biologically active compounds, including pharmaceuticals.

Although it remains to be seen how generally useful the optimal catalyst for C2 functionalization in pentane will be as a tool for synthesis, the authors' platform for catalyst screening and optimization will potentially allow catalyst structures for any given substrate to be tailored as needed. More broadly, Liao and colleagues' research represents an important step towards achieving high selectivity in catalytic C–H functionalization, even in the most challenging contexts. Reactions of this type would allow organic molecules to be groomed, pruned and shaped with bonsai-like precision, readying them for an array of potential applications. ■

**Kin S. Yang and Keary M. Engle** are in the Department of Chemistry, Scripps Research Institute, La Jolla, California 92037, USA. e-mail: keary@scripps.edu

1. Liao, K., Negretti, S., Musaev, D. G., Bacsa, J. & Davies, H. M. L. *Nature* **533**, 230–234 (2016).
2. Labinger, J. A. & Bercaw, J. E. *Nature* **417**, 507–514 (2002).
3. Arndtsen, B. A., Bergman, R. G., Mobley, T. A. & Peterson, T. H. *Acc. Chem. Res.* **28**, 154–162 (1995).
4. Brückl, T., Baxter, R. D., Ishihara, Y. & Baran, P. S. *Acc. Chem. Res.* **45**, 826–839 (2012).
5. Neufeldt, S. R. & Sanford, M. S. *Acc. Chem. Res.* **45**, 936–946 (2012).
6. Davies, H. M. L. & Beckwith, R. E. J. *Chem. Rev.* **103**, 2861–2903 (2003).
7. Davies, H. M. L. & Morton, D. *Chem. Soc. Rev.* **40**, 1857–1869 (2011).

## ATMOSPHERIC SCIENCE

# Ancient air caught by shooting stars

**Ashes of ancient meteors recovered from a 2.7-billion-year-old lake bed imply that the upper atmosphere was rich in oxygen at a time when all other evidence implies that the atmosphere was oxygen-free. SEE LETTER P.235**

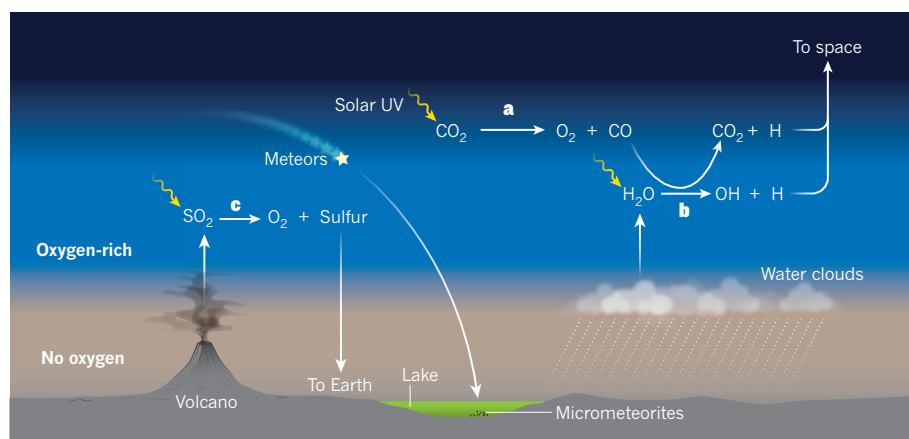
KEVIN ZAHNLE & ROGER BUICK

It is a truth almost<sup>1</sup> universally acknowledged that Earth's atmosphere before about 2.5 billion years ago had little or no free oxygen. The classic argument for anoxia on ancient Earth is that a distinct change occurred in the oxidation state of many surface rocks and minerals around the end of the Archaean eon<sup>2</sup> (which lasted from 4 billion to 2.5 billion years ago). A more recent argument is that a sudden, permanent change in the relative abundances

of rare sulfur isotopes preserved in sediments also occurred at that time — a change that can be linked to differences in sulfur's atmospheric chemistry in the presence or absence of oxygen<sup>3</sup>. These arguments are strong. It therefore comes as a surprise that melted meteor fragments recovered from Archaean limestone indicate that the contemporaneous atmosphere above 75 kilometres was highly oxidized, as reported by Tomkins *et al.*<sup>4</sup> on page 235.

The authors recovered 60 micrometeorites from 2.7-billion-year-old limestone





**Figure 1 | Possible mechanisms of oxygen enrichment in the Archaean atmosphere.** Tomkins *et al.*<sup>4</sup> conclude that micrometeorites found in sedimentary limestone deposited 2.7 billion years ago were oxidized by their passage through an oxygen-rich upper atmosphere at a time when the lower atmosphere was anoxic. **a**, Oxygen could have formed in the upper atmosphere if carbon dioxide was split by solar ultraviolet radiation to form oxygen and carbon monoxide. However, the CO would have to have been removed to enrich the upper atmosphere in oxygen. **b**, Water vapour (most of which is likely to have condensed in the lower atmosphere) could have split to form OH radicals and hydrogen atoms, so that the OH could react with CO to regenerate CO<sub>2</sub>. The hydrogen would have escaped to space, leaving oxygen behind. **c**, Alternatively, if sulfur dioxide from volcanoes was split by solar UV to form oxygen and sulfur particles, then the particles could have fallen to Earth's surface, leaving oxygen behind.

in the Pilbara region of Western Australia. Micrometeorites are the surviving bits of meteors that were too small to burn up as shooting stars in the atmosphere, and are typically tens of micrometres in diameter. All but one of the Pilbara micrometeorites were originally sand-sized grains of iron and nickel alloy. Of the 11 studied in detail, 9 are composed of an oxidized mineral called magnetite (Fe<sub>3</sub>O<sub>4</sub>) and retain a distinctive morphology that indicates fast cooling. The other two retain some of the original metal and wüstite (FeO), an iron oxide that occurs in meteorite fusion crusts. Nearly as remarkable as their existence is that there is nothing otherwise remarkable about them — they look like, and are as oxidized as, the iron micrometeorites that fall to Earth today<sup>5</sup>.

Tomkins *et al.* argue that the micrometeorites look modern because the air above 75 km was roughly as oxidized during the Archaean as it is today, and back this up using a model of meteor physics and chemistry tuned for modern Earth. The model shows that the air in that region needed to be oxygen-rich for it to oxidize all the iron to magnetite as it slowed the meteor's flight. If, however, the chemical reaction continued for a little longer than was modelled, it might have sufficed for the air to be less oxid. Carbon dioxide could have acted as an alternative oxidant (see Extended Data Fig. 5 of the paper<sup>4</sup>), although the kinetics for oxidation with CO<sub>2</sub> are less favourable than with oxygen.

The idea that an oxygen-rich upper atmosphere sat on top of an anoxic lower atmosphere poses a serious challenge to atmospheric modellers. Models predict<sup>6</sup> that oxygen can be abundant at extremely high altitudes in otherwise anoxic atmospheres, but because the oxygen comes from CO<sub>2</sub> that was split by sunlight (photolysis), it is balanced by a stoichiometric

complement of carbon monoxide (Fig. 1). Overall, the resulting gas mixture would be no more oxidizing than CO<sub>2</sub> itself.

To create a local superabundance of oxygen from CO<sub>2</sub>, the CO must be preferentially removed. There is no obvious way to do this. Some other molecule is therefore required that can be split into oxygen and a chemically reduced species that can be easily removed. One possible candidate is sulfur dioxide (SO<sub>2</sub>) from volcanoes. This gas can be split by sunlight into oxygen and elemental sulfur, which can condense to form particles that fall to Earth, leaving oxygen behind. Isotopes in sedimentary rock indicate that elemental sulfur did fall from the skies during the Archaean<sup>7</sup>, which makes SO<sub>2</sub> an attractive candidate.

The other obvious candidate is water vapour, which can be split by sunlight to free hydrogen atoms that escape to space, leaving oxygen behind. Water can do double duty here: the hydroxyl (OH) radicals generated by water photolysis react with CO to put CO<sub>2</sub> back together again, freeing more hydrogen atoms. If the hydrogen atoms escape quickly enough, then the top of the atmosphere can become rich in oxygen. This is why Mars has more oxygen than CO in its atmosphere.

But today, water is cold-trapped in Earth's lower atmosphere and the upper atmosphere is very dry. For water vapour to have reached the upper atmosphere during the Archaean, either the cold trap must have been warmer than it is today, or the atmosphere must have been thinner, which would have made the cold trap less effective<sup>8</sup>. Evidence<sup>9</sup> that atmospheric pressure was less than half of what it is today — and perhaps much less — has been found in rocks of almost identical age to those hosting the oxidized micrometeorites. The micrometeorites

might, therefore, be evidence of a thinner atmosphere.

The Pilbara micrometeorites are not the only preserved probes of the Archaean atmosphere. Several thick beds of impact-generated spherules — rounded bodies formed from the molten ejecta of meteorite impacts — provide hints about the atmosphere they fell through<sup>10</sup>. The Archaean spherule beds are analogous to the thinner spherule layers broadcast worldwide by the Chicxulub impact that killed the dinosaurs. One distinction between the Chicxulub and the Archaean spherules is that the latter were formed under markedly more reducing conditions, which has been interpreted to mean that the Archaean atmosphere contained no more than 0.01% oxygen<sup>10</sup>. The spherules probably last reacted with air when they re-entered the atmosphere as meteors — typically reaching a height of between 30 and 50 km, judging from the spherules' size (about a millimetre in diameter). In other words, the air last sampled by the spherules would have been between the air at Earth's surface and that sampled by the micrometeorites.

It is remarkable that objects as small as the micrometeorites survived intact for 2.7 billion years. The survival of wüstite is particularly unusual — this mineral is not normally seen near Earth's surface — and is crucial to the authors' interpretation of these minuscule objects as being extraterrestrial. However, the micrometeorites were deposited in a highly unusual environment: the Tumbiana Formation.

This rock formation was once a system of lakes, and the lake in which the micrometeorites were found was highly alkaline<sup>11</sup>, as indicated by its extremely high abundance of heavy nitrogen isotopes<sup>12</sup>. Wüstite has low solubility under such pH conditions<sup>13</sup>, and would have been especially insoluble if the bottom waters of the lake and the pore waters in the buried sediments were anoxic. Such conditions are rarely encountered in the geological record, which means that the Pilbara micrometeorites might be a one-off discovery. This would be unfortunate, because the structure, pressure and vertical composition of the ancient atmosphere are fiendishly difficult things to determine. But one can wish upon a shooting star. ■

**Kevin Zahnle** is in the Space Science Division, NASA Ames Research Centre, Moffett Field, California 94035-1000, USA. **Roger Buick** is in the Department of Earth & Space Sciences, and in the Astrobiology Program, University of Washington, Seattle, Washington 98195-1310, USA.

e-mails: kevin.j.zahnle@nasa.gov;  
buick@ess.washington.edu

1. Ohmoto, H. *Geochem. News* **93**, 12–13, 26–27 (1997).
2. Holland, H. D. *Geochem. News* **100**, 20–22 (1999).
3. Farquhar, J., Bao, H. & Thiemens, M. *Science* **289**, 756–758 (2000).
4. Tomkins, A. G. *et al. Nature* **533**, 235–238 (2016).

5. Genge, M. J., Engrand, C., Gounelle, M. & Taylor, S. *Meteor. Planet. Sci.* **43**, 497–515 (2008).
6. Zahnle, K., Claire, M. & Catling, D. *Geobiology* **4**, 271–283 (2006).
7. Pavlov, A. A. & Kasting, J. F. *Astrobiology* **2**, 27–41 (2002).
8. Wordsworth, R. & Pierrehumbert, R. *Astrophys. J.* **785**, L20 (2014).
9. Som, S. M. *et al.* *Nature Geosci.* <http://dx.doi.org/10.1038/ngeo2713> (2016).
10. Krull-Davatzes, A. E., Byerly, G. R. & Lowe, D. R. *Earth Planet. Sci. Lett.* **296**, 319–328 (2010).
11. Awramik, S. M. & Buchheim, H. P. *Precamb. Res.* **174**, 215–240 (2009).
12. Stüeken, E. E., Buick, R. & Schauer, A. J. *Earth Planet. Sci. Lett.* **411**, 1–10 (2015).
13. Jang, J. H. & Brantley, S. L. *Environ. Sci. Technol.* **43**, 1086–1090 (2009).

## QUANTUM-MATTER PHYSICS

# Quasiparticles on a collision course

**Emergent quanta of momentum and charge, called quasiparticles, govern many of the properties of materials. The development of a quasiparticle collider promises to reveal fundamental insights into these peculiar entities. SEE LETTER P.225**

DIRK VAN DER MAREL

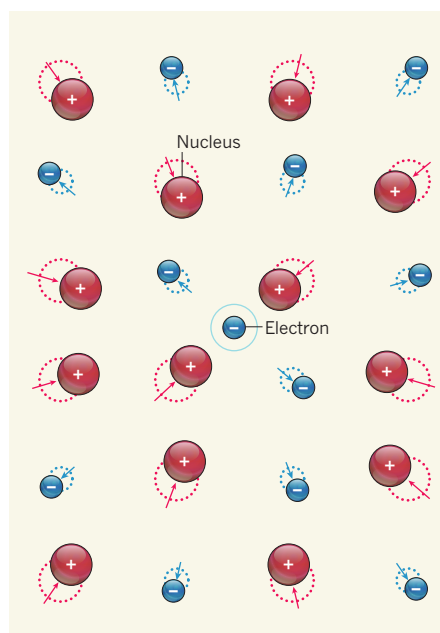
Electrons and atomic nuclei in solids are bound together by their electric charges. If an electron is moved to a new position inside a metal, then the other electrons and nuclei respond by shifting their own positions. An electron that is accompanied by this response of the surrounding electrons and nuclei is an example of a quasiparticle (Fig. 1). It would be fascinating to prepare and manipulate the trajectories of quasiparticles to make them collide and then to study the effect of the collision, similar to experiments in a particle accelerator. On page 225 of this issue, Langer *et al.*<sup>1</sup> report the realization of such an experiment.

The authors studied an electrical insulator, tungsten diselenide (WSe<sub>2</sub>). They generated pairs of quasiparticles in the material, one negatively charged and the other positively charged, using an ultrashort light pulse (10–100 femtoseconds in duration; 1 fs is 10<sup>–15</sup> seconds). The light pulse's energy, intensity and duration were precisely adjusted so that the initial distance of the quasiparticles from each other, and their relative speeds, were well defined.

Langer and colleagues then launched the quasiparticles along a linear track. This track was created with the help of the electric field from a second light pulse; the field strength, duration and oscillation period of the light pulse were adjusted to direct the quasiparticles into a head-on collision. The collision caused mutual annihilation of the quasiparticles and the emission of a photon, which the authors detected. The experiment is therefore similar to studies of electron–positron annihilation in high-energy particle accelerators (positrons are the antiparticles of electrons, which means that they have opposite charge and equal mass to an electron).

The researchers can tune the conditions of their system in many ways by adjusting the

forementioned experimental parameters and the time interval between the generation of the pulses and their detection. They particularly examined the effect of the electrical (Coulomb) interaction between the two oppositely charged quasiparticles. Under stable conditions, this interaction would bind the quasiparticles into a neutral composite particle called an exciton. Excitons are another example of a collective state that can exist in solids,



**Figure 1 | An emergent quasiparticle.** When an electron moves to a new position inside a metal, the other electrons and atomic nuclei shift their own positions in response (empty circles show original positions of the surrounding electrons and nuclei; electrons are negatively charged, nuclei are positively charged). The combination of the electron and the movement of the surrounding electrons and nuclei is an example of a quasiparticle — a collective phenomenon that behaves like a single particle. Langer *et al.*<sup>1</sup> describe a system that allows quasiparticle collisions to be studied.

somewhat like positronium atoms (which form from one electron and one positron). The authors obtained material-specific information such as the exciton binding energy, and observed an enhancement of the collisional cross-section (a quantity that governs the rate at which the quasiparticles collide) as a result of the Coulomb force between the oppositely charged quasiparticles.

The beauty of Langer and colleagues' experimental toolkit is that it might finally allow quasiparticles and their mutual interactions to be studied in the materials in which they arise. The negative and positive quasiparticles in the authors' experiment are similar to electrons and positrons in a vacuum, but a rich variety of unconventional quasiparticles could also be studied, for which no equivalent elementary particles are known. For example, when an electron is introduced into an insulating transition-metal oxide such as strontium titanate (SrTiO<sub>3</sub>), the electron slightly attracts the positive ions in the material (Ti<sup>4+</sup> and Sr<sup>3+</sup>), but slightly repels the negative oxide ions (O<sup>2–</sup>). When the electron moves around the compound's lattice, the ionic displacements move with the electron. The resulting object — the electron plus the co-moving lattice distortion — is called a polaron<sup>2,3</sup>. Its properties and behaviour are different from those of an electron; for example, its mass is typically two or three times higher.

Quasiparticles that are even more bizarre emerge in two-dimensional gases of interacting electrons in a strong magnetic field. The charge on these quasiparticles is a fraction of that for an electron: it can be one-third (the same as for elementary particles called quarks), one-fifth, one-seventh, or smaller<sup>4,5</sup>.

When a magnetic field is applied to certain superconductors, peculiar topological states known as vortices appear, equivalent to tubes of magnetic flux. Vortices and antivortices form spontaneously<sup>6,7</sup> in 2D superconductors, but it might also be possible to generate them using light pulses. This would open the way to studies of their interactions using Langer and colleagues' approach, including the annihilation of vortex–antivortex pairs.

Quasiparticles are not only of academic interest — they also determine many of the properties and functionalities of materials, such as electrical resistivity, heat capacity and magnetism. There are thus many reasons to study quasiparticles in the materials in which they are manifested. Langer and co-workers have provided a fresh strategy with which condensed-matter physicists can tackle such studies. This promises fundamental insights,



but also offers ways to control and handle the quasiparticles characteristic of the various states of matter that can be realized in solids.

That said, only a few experimental facilities will have the combination of technologies required to study quasiparticles that have fractional charges, or vortex–antivortex annihilation in two dimensions. But for those that do, Langer and colleagues' approach can be readily applied to investigate the properties of polarons in strontium titanate or other transition-metal oxides, or the 'heavy electrons' that occur in several materials owing to the coupling of mobile electrons to fluctuations of

magnetic polarization<sup>8–10</sup>. According to some schools of thought, the quasiparticle concept does not apply in certain materials or under special conditions<sup>11</sup>. Collision experiments might therefore help to identify the boundaries of the quasiparticle concept. ■

**Dirk van der Marel** is in the Department of Quantum Matter Physics, University of Geneva, CH-1211 Geneva 4, Switzerland. e-mail: dirk.vandermarel@unige.ch

1. Langer, F. et al. *Nature* **533**, 225–229 (2016).
2. Eagles, D. M., Georgiev, M. & Petrova, P. C. *Phys. Rev. B* **54**, 22–25 (1996).

3. van Mechelen, J. L. M. et al. *Phys. Rev. Lett.* **100**, 226403 (2008).
4. Tsui, D. C., Stormer, H. L. & Gossard, L. A. C. *Phys. Rev. Lett.* **48**, 1559–1562 (1982).
5. Laughlin, R. B. *Phys. Rev. Lett.* **50**, 1395–1398 (1983).
6. Berezinskii, V. L. *Sov. Phys. JETP* **32**, 493–500 (1971).
7. Kosterlitz, J. M. & Thouless, D. J. *J. Phys. C* **6**, 1181–1203 (1973).
8. Stewart, G. R., Fisk, Z., Willis, J. O. & Smith, J. L. *Phys. Rev. Lett.* **52**, 679–682 (1984).
9. De Visser, A., Franse, J. J. M., Menovsky, A. & Palstra, T. T. M. *Physica B+C* **127**, 442–447 (1984).
10. Mackenzie, A. P. & Maeno, Y. *Rev. Mod. Phys.* **75**, 657–712 (2003).
11. van der Marel, D. et al. *Nature* **425**, 271–274 (2003).

## CELESTIAL MECHANICS

# Fresh solutions to the four-body problem

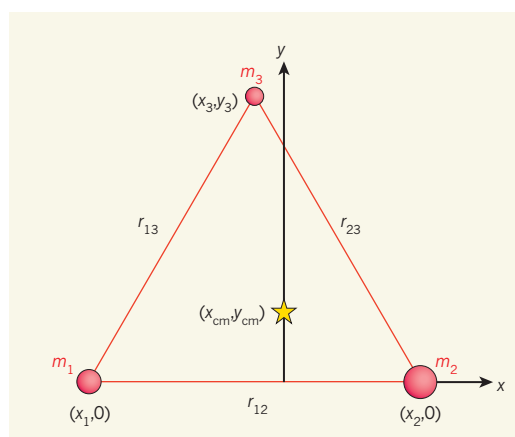
**Describing the motion of three or more bodies under the influence of gravity is one of the toughest problems in astronomy. The report of solutions to a large subclass of the four-body problem is truly remarkable.**

DOUGLAS P. HAMILTON

The study of the orbital motions of bodies that are subject to their mutual gravitational attractions is crucial for understanding the movements of moons, planets and stars, and for navigating spacecraft to distant planets. The central problem is to determine the motions of  $n$  point masses interacting through gravitational forces that vary with the inverse square of their separation distances. This  $n$ -body problem is famous among astronomers and mathematicians, and is known to have no general analytical solution (that is, no solution that can be written down in terms of simple mathematical functions). Nevertheless, specific solutions have been eagerly sought and occasionally discovered. Writing in *Celestial Mechanics and Dynamical Astronomy*, Érdi and Czirják<sup>1</sup> report analytical solutions for a broad class of four-body configurations.

Isaac Newton solved the two-body problem in his 1687 masterwork, the *Principia*, but the three-body problem proved surprisingly complex and occupied many distinguished mathematicians over the next two centuries. Leonhard Euler and Joseph-Louis Lagrange found all analytical solutions to an important subclass of the three-body problem known as central configurations, but work by Heinrich Bruns and by Henri Poincaré in the late 1880s showed that a general arrangement of three

or more bodies admits no analytical solution. Although the set of all possible central configurations of four bodies remains unknown, Érdi and Czirják have taken a large stride forward by solving all of those in which two of the



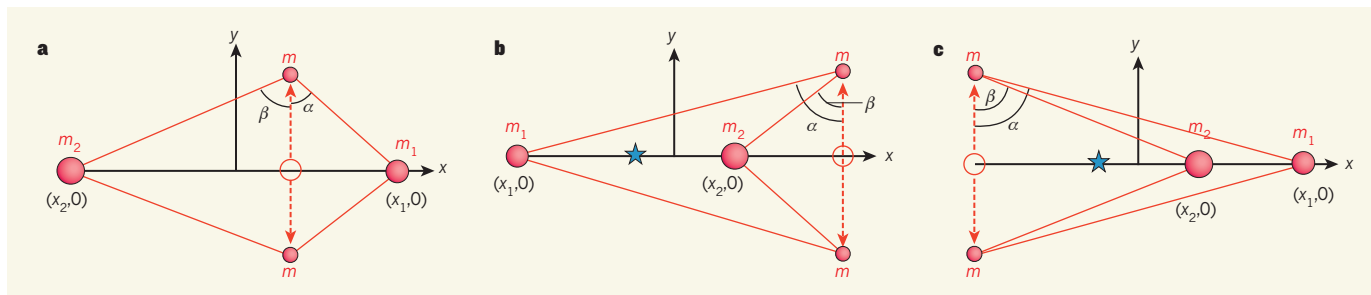
**Figure 1 | A subclass of the three-body problem.** The motions of three bodies with masses  $m_1$ ,  $m_2$  and  $m_3$  under the influence of gravitational forces can be described analytically — that is, in terms of simple mathematical functions — for the special case in which the bodies are placed at the vertices of an equilateral triangle. Proof of this involves considering the accelerations in the  $y$  direction of two masses placed on the  $x$  axis. It emerges that the bodies orbit in such a way that the triangle rotates, expands or shrinks, and always remains in the  $xy$  plane. The yellow star represents the centre of mass of the three-body system. Distances between the masses are represented by the symbol  $r$ , with subscripts representing the masses; coordinates for masses and for the centres of mass are given as  $(x, y)$  pairs.

bodies lie along an axis of symmetry.

In central configurations, each body must be subject to an acceleration directed towards the centre of mass of the system with a magnitude that is proportional to its distance from the centre of mass. All orbits of two bodies are central configurations, in which the objects each orbit their common centre of mass along ellipses that have identical shapes and orbital periods. Euler's solutions for linear arrangements of three bodies are also central configurations, as are Lagrange's solutions in which the three masses are placed at the vertices of an equilateral triangle. In the latter system, Lagrange showed that the vertices of the triangle can move in such a way as to preserve relative distances between the masses; the triangle can rotate around the centre of mass, expand or shrink, but must remain in its initial plane.

It is easy to show that equilateral triangles are the only possible planar central configuration of a three-body system by placing two of the masses ( $m_1$  and  $m_2$ ) on an  $x$  axis, and considering their accelerations in the perpendicular  $y$  direction (Fig. 1). The  $y$  acceleration on  $m_1$  is due solely to the gravity of the third mass ( $m_3$ ) and equals  $Gm_3y_3/r_{13}^3$  — where  $G$  is the gravitational constant,  $y_3$  is the coordinate of  $m_3$  along the  $y$  axis and  $r_{13}$  is the distance between  $m_1$  and  $m_3$ . The corresponding  $y$  acceleration on  $m_2$  is  $Gm_3y_3/r_{23}^3$ . According to the definition of central configurations, these accelerations must separately equal  $\lambda y_{cm}$  (where  $y_{cm}$  is the  $y$  coordinate of the centre of mass and  $\lambda$  is the common proportionality constant). By cancelling like terms in the two  $y$  accelerations, it immediately becomes apparent that  $r_{13}$  must be the same as  $r_{23}$ .

If the symmetry of the equilateral-triangle system is then exploited by choosing a new  $x$  axis to run along the line connecting  $m_1$  and  $m_3$ , repeating the above argument shows that  $r_{12}$  must also be the same as  $r_{23}$ , and thus all three sides of the triangle must be equal in length. This proof extrapolates directly to four bodies: the only fully three-dimensional central configurations for



**Figure 2 | A subclass of the four-body problem.** Érdi and Czirják<sup>1</sup> solved a subclass of the four-body problem derived from systems of three masses initially located on the  $x$  axis. **a–c**, In each case, two masses ( $m_1$  and  $m_2$ ) are left on the  $x$  axis; the third mass (original position shown with open circle) is split into two equal halves, each of mass  $m$ , which are moved symmetrically in the  $y$  direction (dashed arrows). The centre of mass of each system is at

$(0,0)$  and the masses form convex (**a**) or concave (**b** and **c**) polygons when connected. The centre of mass excluding  $m_2$  is shown with a blue star in **b** and **c**; its location inside or outside the polygon distinguishes the two cases. The positions of  $m_1$  and  $m_2$  were defined using rectangular coordinates, whereas those of the two off-axis masses were fixed by the angles  $\alpha$  and  $\beta$ .

four bodies are those in which the masses are placed on the vertices of a tetrahedron such that all distances between the masses are equal.

But two-dimensional central solutions of four bodies are much more difficult to identify. In their seminal work, Érdi and Czirják find three examples of such solutions, which can be visualized by considering a system of three masses distributed on a line. Each solution is found by splitting one of the masses into equal halves and moving the fragments up and down so that the resulting distribution of four masses is symmetric about the  $x$  axis (Fig. 2). The four-sided polygon formed by connecting the on-axis masses to the off-axis masses is convex when the central mass is split (Fig. 2a), and concave when one of the other masses is split (Fig. 2b,c). Érdi and Czirják's two concave cases differ according to whether the centre of mass of the system excluding  $m_2$  is enclosed by the polygon or not.

The next step would normally be to specify the masses and then to seek all arrangements of those masses that satisfy the conditions for a central configuration. Érdi and Czirják, however, chose to tackle the inverse problem: given the positions of the bodies, they computed the masses that make the configuration central. And, rather than working with rectangular coordinates for the two off-axis masses, the authors chose to recast the problem in terms of a pair of angles that fix the position of those masses relative to the ones on the  $x$  axis (Fig. 2). These are both inspired choices that make the problem analytically tractable. If one or more of the four masses is set to zero, the angles take on values that are consistent with straight lines and equilateral triangles; in this way the four-body edifice of Érdi and Czirják's work is rooted in the three-body bedrock of Euler and Lagrange.

Central configurations are dynamic equilibria that can be stable (such as a ball at the bottom of a smooth bowl) or unstable (as for a ball perched atop a round hill). Euler's straight-line configurations are all unstable so that, like the ball on the hill, the configuration cannot persist when tweaked. Lagrange's equilateral-triangle solutions are stable if

one of the three masses contains more than about 96% of the total mass of the system, but unstable if the mass is more evenly distributed. Thus, Lagrangian configurations for a system that incorporates the Sun, Jupiter and a suitably placed asteroid are stable, as would be those for Earth, the Moon and a modestly sized future space station. By contrast, Pluto's massive moon Charon prevents any central configurations involving these bodies and a smaller moon from being stable. Whether the new four-body central configurations are stable is an interesting, unexplored question and is an inviting direction for future research.

Érdi and Czirják's solution to a large subclass of the central four-body problem is a major advance that encompasses and greatly extends many previous four-body results, including: arrangements of four<sup>2</sup> and three<sup>3</sup> identical masses; kite-shaped configurations of diagonally opposite pairs of equal masses<sup>4</sup>; and the limiting case of three bodies plus a massless test particle<sup>5</sup>. Just as three-body configurations serve as limiting cases for Érdi

and Czirják's four-body configurations, the authors' solutions could, in turn, be used as limiting cases for ambitious future extensions of the  $n$ -body problem: perhaps three masses along a line plus two symmetrically placed equal masses; a test particle plus planar configurations of the type considered in the present work; or even planar arrangements of four different masses. ■

**Douglas P. Hamilton** is in the Department of Astronomy, University of Maryland, College Park, Maryland 20742-2421, USA.  
e-mail: dphamil@astro.umd.edu

1. Érdi, B. & Czirják, Z. *Celest. Mech. Dyn. Astron.* **125**, 33–70 (2016).
2. Albouy, A. *Contemp. Math.* **198**, 131–135 (1996).
3. Long, Y. & Sun, S. *Arch. Ration. Mech. Anal.* **162**, 25–44 (2002).
4. Alvarez-Ramírez, M. & Llibre, J. *Appl. Math. Comput.* **219**, 5996–6001 (2013).
5. Piña, E. & Lonngi, P. *Celest. Mech. Dyn. Astron.* **108**, 73–93 (2010).

This article was published online on 4 May 2016.

## NEUROBIOLOGY

# Wired for sex

**Analysis of a sensory neural circuit in the roundworm *Caenorhabditis elegans* reveals that its wiring is sex-specific, and arises through the elimination of connections that are originally formed in both sexes. SEE ARTICLE P.206**

**DOUGLAS S. PORTMAN**

If presented with a human brain, even the most meticulous neuroanatomist would be hard-pressed to identify the sex of its former owner. There are clearly male–female differences in some brain regions, but these can be subtle and variable, and their causes and consequences remain largely unclear. Over the past five years, work in several organisms<sup>1–3</sup> has suggested that altered neural connectivity

between brain regions might be a hallmark of male–female differences. On page 206 of this issue, Oren-Suissa *et al.*<sup>4</sup> provide clear evidence for sex differences in neural wiring in the roundworm *Caenorhabditis elegans*. Moreover, they report that these differences arise through sex-specific eradication of neural connections and are controlled by the genetic sex of the nervous system itself.

In *C. elegans*, males are males, but females are properly called hermaphrodites. They



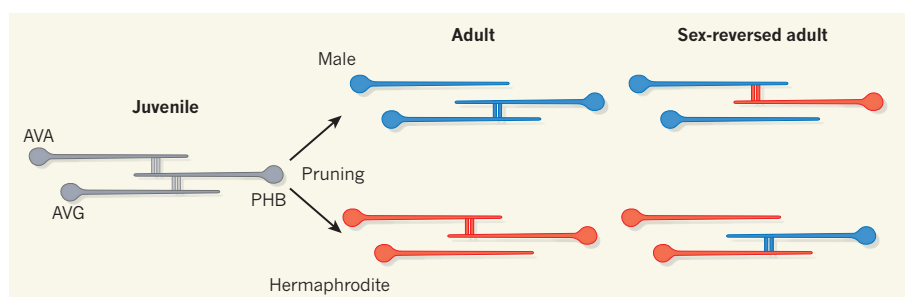
produce a reserve of sperm as juveniles and can reproduce either by self-fertilization or by mating with a male. The nervous systems of both sexes have been studied in extraordinary detail. An adult hermaphrodite has precisely 302 neurons, and the complete set of connections between these neurons — the connectome — was mapped<sup>5</sup> in the 1980s. These connections, called synapses, wire neurons up into circuits and are crucial for brain function. The complete hermaphrodite connectome remains the only one of its kind, and has provided unparalleled information about neural circuits and their control of behaviour.

Oren-Suissa *et al.* began by comparing the hermaphrodite's wiring diagram with the connectome of the adult male's tail<sup>6</sup>, which was determined in 2012. The authors focused on a circuit of interconnected neurons that receives and interprets input from a tiny sensory organ in the tail called the phasmid. Unexpectedly, they identified seven synapses present in one sex but not the other, a phenomenon known as sexual dimorphism. For example, although two neurons called PHB and AVA are adjacent in both sexes, a synaptic connection between them seemed to exist only in hermaphrodites. In males, PHB connected instead to the nearby neuron AVG (Fig. 1).

The investigators used genetic tools to fluorescently label these seven synapses, allowing the connections to be visualized in live animals. After confirming that the connections really were sexually dimorphic, Oren-Suissa and colleagues asked how sex influenced synaptic patterning. A key insight came from examining sexually immature animals. At five of the synapses, juvenile animals had both male- and hermaphrodite-specific connections, indicating that both sexes initially adopt a sexually intermediate state. Sexual dimorphism arises through the subsequent elimination of specific connections — a process called synaptic pruning.

Although synaptic pruning is not well described in *C. elegans*, it is essential in the development of the mammalian nervous system, enabling changes in connectivity that are thought to underlie learning, memory and cognitive maturation. Cells called microglia, which specialize in synaptic pruning, have been implicated in the generation of sex differences in the mammalian brain<sup>7</sup>. Thus, pruning could have an evolutionarily conserved role in shaping sex-specific differences in circuit connectivity. Furthermore, disrupted synaptic pruning has been suggested to contribute to neuropsychiatric disorders that show sex biases, such as autism<sup>8</sup> and schizophrenia<sup>9</sup>. Perhaps differences in the pruning of particular connections can make one sex — in these cases, males — more susceptible to disease.

Oren-Suissa *et al.* investigated the underlying mechanisms that link sex to the pattern of pruning. In *C. elegans*, an individual's sex is determined by the number of X chromosomes



**Figure 1 | Lost in maturation.** Oren-Suissa *et al.*<sup>4</sup> identify seven synaptic connections between neurons in the roundworm *Caenorhabditis elegans* that are sex-specific, including those between neurons called PHB and AVA, and those between PHB and AVG. In juvenile animals, these synapses are found in both sexes — male and hermaphrodite. As worms mature, synapses are eliminated in a manner that depends on the genetic sex of each neuron. In males, the connection between PHB and AVA is lost. This connection is maintained in hermaphrodites, whereas that between PHB and AVG is lost. Genetically switching the sex of PHB is sufficient to reverse this pattern of connections.

it carries. Reversing this genetic signal in the nervous system — or even in individual neurons — has been shown to alter circuit physiology and behaviour<sup>10</sup>. Using this approach, the researchers found that the genetic sex of individual neurons can also determine whether a particular synapse is pruned. For example, genetically 'masculinizing' the PHB neuron of a juvenile hermaphrodite usually led to elimination of the PHB–AVA connection. Moreover, the PHB–AVG synapse, which is typically pruned in hermaphrodites, was retained (Fig. 1). Thus, the stability of particular synapses is linked to genetic sex.

At first glance, this role for genetic sex might seem inconsequential for mammals, in which information about sexual state is broadcast throughout the body by hormones — such as testosterone or oestrogen — released by the gonads. However, sexual differentiation of the mammalian brain also depends on its own genetic sex<sup>11</sup>. The details of the differences caused by this effect remain largely mysterious, and studies in invertebrates may provide insight into its workings.

Oren-Suissa *et al.* next considered the consequences of sex-specific pruning on behaviour. In hermaphrodites, pruning generates a phasmid circuit that regulates avoidance of noxious chemicals. In males, however, pruning links the phasmid to neurons that signal the presence of mates. The authors therefore wondered whether the phasmid circuit might be repurposed for male sexual behaviour. Although they did not test the behaviour of males that had a feminized phasmid circuit, they did find that disrupting the phasmid by disabling the PHB neuron compromised males' ability to mate.

This study leaves some questions unexamined, and raises fascinating new ones. It is easy to understand why juvenile males benefit from avoiding noxious chemicals, providing a rationale for hermaphrodite-like connections in an immature male. But it is less obvious why young hermaphrodites would build male-like connections, only to prune

them later. It also remains unclear how a given neuron 'knows' which of its many synapses to prune. In mammals, pruning is often a consequence of decreased synaptic activity; perhaps this phenomenon plays a part in *C. elegans*, too. Finally, does genetic sex also prune connections in the worm's head, which harbours circuits that control more-complex decision-making? And what might the consequences of this be?

Oren-Suissa and colleagues' study provides key support for the emerging idea that biological sex modulates the structure and function of neural circuits. In more-complex animals, sex differences in pruning might modulate functional connectivity, disease susceptibility and cognition. Further studies of natural sexual variation in the brain are likely to provide important insight into the workings of neural circuits. They should also shed light on the interactions between social and biological processes, and could perhaps even tell us something about what makes us human. ■

**Douglas S. Portman** is in the Department of Biomedical Genetics and the Center for Neural Development and Disease, University of Rochester, Rochester, New York 14642, USA. e-mail: douglas.portman@rochester.edu

- Kohl, J., Ostrovsky, A. D., Frechter, S. & Jefferis, G. S. X. E. *Cell* **155**, 1610–1623 (2013).
- Bergan, J. F., Ben-Shaul, Y. & Dulac, C. *eLife* **3**, e02743 (2014).
- Ingallhalikar, M. *et al. Proc. Natl Acad. Sci. USA* **111**, 823–828 (2014).
- Oren-Suissa, M., Bayer, E. A. & Hobert, O. *Nature* **533**, 206–211 (2016).
- White, J. G., Southgate, E., Thomson, J. N. & Brenner, S. *Phil. Trans. R. Soc. Lond. B* **314**, 1–340 (1986).
- Jarrell, T. A. *et al. Science* **337**, 437–444 (2012).
- Lenz, K. M., Nugent, B. M., Haliyur, R. & McCarthy, M. M. *J. Neurosci.* **33**, 2761–2772 (2013).
- Frith, C. *Lancet Neurol.* **3**, 577 (2004).
- Feinberg, I. J. *Psychiatr. Res.* **17**, 319–334 (1982–83).
- Ryan, D. A. *et al. Curr. Biol.* **24**, 2509–2517 (2014).
- McCarthy, M. M. & Arnold, A. P. *Nature Neurosci.* **14**, 677–683 (2011).

This article was published online on 4 May 2016.

# Palaeoclimatic insights into forcing and response of monsoon rainfall

Mahyar Mohtadi<sup>1</sup>, Matthias Prange<sup>1</sup> & Stephan Steinke<sup>1</sup>

**Monsoons are the dominant seasonal mode of climate variability in the tropics and are critically important conveyors of atmospheric moisture and energy at a global scale. Predicting monsoons, which have profound impacts on regions that are collectively home to more than 70 per cent of Earth's population, is a challenge that is difficult to overcome by relying on instrumental data from only the past few decades. Palaeoclimatic evidence of monsoon rainfall dynamics across different regions and timescales could help us to understand and predict the sensitivity and response of monsoons to various forcing mechanisms. This evidence suggests that monsoon systems exhibit substantial regional character.**

**R**eliable prediction of summer monsoons is critical to mitigating the often catastrophic consequences of monsoon rainfall anomalies, such as floods and droughts, famine, and economic losses. Yet, despite its pivotal role in the livelihood of billions of people, summer monsoon rainfall remains difficult to predict<sup>1–3</sup>. It is expected that total monsoon precipitation in the Northern and Southern hemispheres will change in opposite directions in the coming decades, owing to differences in hemispheric warming<sup>4</sup>. Regional differences are expected to be large, and some monsoon regions are predicted to receive more rainfall despite a projected weakening of the monsoon circulation<sup>1</sup>. Even though the latest generation of climate models show substantial overall improvements compared to their predecessors<sup>2,5</sup>, the extent and intensity of monsoons are still often under-simulated<sup>2</sup>, with most models showing systematic errors in the seasonal cycle and in intra-seasonal-to-interannual variability<sup>5,6</sup>. Moreover, there is low confidence in projections of future changes in the amount of rainfall for several monsoon domains<sup>1</sup>. For instance, projection of the South Asian (Indian) summer monsoon rainfall is a key challenge for global and regional circulation models<sup>7</sup>. Poor simulations of monsoon dynamics arise mainly from the complexity of these land–ocean–atmosphere coupled systems that interact with nearly all other tropical and extratropical climate phenomena, such as the El Niño–Southern Oscillation (ENSO)<sup>3</sup> and the Hadley and Walker circulations<sup>8</sup> (Box 1).

The monsoon concept has recently changed from a regional to a global one. The new idea of a “global monsoon”<sup>8,9</sup> takes into account a coherent response of all monsoon systems, regardless of regional differences, to changes in global-scale atmospheric circulation patterns forced by the annual cycle of solar radiation and land–air–sea interactions<sup>8</sup>. Modern climatology of monsoon domains (Fig. 1) is well depicted by the concept of a global monsoon<sup>9</sup>, and models are substantially better at representing the global monsoon than the regional monsoons<sup>1</sup>. However, several reconstructions and transient simulations of past monsoon variability emphasize that it has a substantial regional character, rather than being characterized by the common, global dynamics of the monsoon systems at different timescales<sup>10,11</sup>. Study of monsoon evolution beyond the instrumental record is thus essential to improve our understanding of natural and anthropogenic forcing mechanisms of monsoon rainfall and its interactions and teleconnections on various timescales, which will enable us to test and enhance climate models in their ability to represent and predict monsoon dynamics, and to contribute to the climate change discussion.

Within the past decade, palaeoclimate studies have provided new insight into past monsoon variability with unprecedented temporal and

spatial coverage. Here we evaluate the forcing mechanisms of variability in summer monsoon rainfall as evidenced by palaeoclimate research (Fig. 2), and identify several aspects of monsoon dynamics that will improve our understanding of future monsoon response to specific forcing. The scope of this Review is constrained by the available monsoon reconstructions and the timeliness of climate model simulations, and excludes oceanic monsoon domains, highly uncertain forcing mechanisms and tectonic forcing.

The palaeoclimatic evidence shows that monsoon dynamics are strongly shaped by large-scale meridional temperature gradients and the related position of the intertropical convergence zone<sup>12</sup> (ITCZ; Box 1). However, study of past monsoons also reveals that these temperature gradients are sensitive to many types of forcing, the influence of which seems to vary in time and space. Until climate models converge on the simulation of monsoon forcings, a robust projection of future monsoons will remain elusive<sup>13</sup>. In this Review, we discuss the main forcings of monsoon variability and their uncertainties, and argue that a coordinated effort to quantify past variations in meridional temperature gradients, including site-specific monsoon reconstructions, will provide a crucial test bed for model improvement.

## Orbital forcing

Changes in the tilt of Earth's axis (obliquity) as well as axial and apsidal precession (wobble of Earth's axis and rotation of Earth's elliptical orbit over time, respectively) modulate the temporal and spatial distribution of insolation. Precession, with periods of about 19 kyr and 23 kyr, affects the seasonal cycle of incoming solar radiation and its hemispheric distribution, and is therefore considered a major control on changes in monsoon intensity<sup>14</sup>. A prominent example of this control was during the early-to-mid Holocene, when higher-than-today Northern Hemisphere summer insolation rendered the North African monsoon strong enough to turn the Sahara desert into a steppe or savannah landscape<sup>15,16</sup>.

Generally, monsoon rainfall and seasonality (that is, the amplitude of the annual rainfall cycle) are enhanced in the Northern Hemisphere and reduced in the Southern Hemisphere during a precession minimum (when the Northern Hemisphere summer solstice occurs at perihelion—the point in the orbit of the Earth at which it is nearest to the Sun) and vice versa during a precession maximum (when the Northern Hemisphere summer solstice occurs at aphelion—the point in the orbit of the Earth at which it is farthest from the Sun); see Fig. 3. Substantial support for this view is provided by more than a decade of research on stable oxygen isotope ( $\delta^{18}\text{O}$ ) records of high-resolution and absolute-dated cave

<sup>1</sup>MARUM—Center for Marine Environmental Sciences, University of Bremen, 28359 Bremen, Germany.



## BOX 1

## Key characteristics of monsoon systems and related climate phenomena

**Intertropical convergence zone (ITCZ).** The ITCZ is a tropical belt of maximum precipitation that results from deep convection that migrates seasonally towards the warming hemisphere<sup>37</sup>. Over land, it moves back and forth across the equator following the zenith point of the Sun. The zonal-mean position of the ITCZ is associated with the rising branch of the global Hadley cell.

**Tropical and subtropical monsoons.** In tropical monsoons, precipitation occurs almost entirely within an ITCZ that is seasonally displaced in the presence of cross-equatorial pressure gradients. The onset of tropical monsoons correlates with a distinct change in the wind shear, such as in the South Asian monsoon<sup>28,43,99</sup>. In subtropical monsoons, precipitation is mainly controlled by the position and migration of the subtropical highs and frontal systems. For the East Asian monsoon, the jet stream and large-scale topography are also crucial to the seasonal evolution of rainfall<sup>28</sup>.

**Dynamic and thermodynamic mechanisms.** Monsoon rainfall is affected by dynamic and thermodynamic mechanisms. Dynamic mechanisms refer to changes in winds with unchanged water-vapour concentration in the air. Thermodynamic mechanisms refer to changes in water-vapour concentration with unchanged winds. Thermodynamic mechanisms dominate the hemispherically antisymmetric, annual-mean precipitation response to precession in the absence of land–sea contrasts<sup>24</sup>.

**Atlantic Multidecadal Oscillation (AMO).** The AMO is a mode of natural climate variability characterized by a coherent pattern of variability in basin-wide North Atlantic sea surface temperature (SST) with a period of 60–80 years, and is associated with multidecadal variations in the strength of the Atlantic overturning circulation<sup>100</sup>.

**Pacific Decadal Oscillation (PDO).** The PDO is the leading empirical orthogonal function of monthly SST anomalies over the North Pacific, and is closely related to the strength of the wintertime Aleutian low-pressure system. It is characterized by SST anomalies with a period of 20–30 years and opposite signs in the western extratropical and eastern Pacific.

**El Niño–Southern Oscillation (ENSO).** The ENSO is a naturally occurring, interannual fluctuation in equatorial Pacific SST, with a warm (El Niño) and a cool (La Niña) phase. Its atmospheric component, the Southern Oscillation, is measured by the sea-level pressure difference between Darwin, Australia and Tahiti, French Polynesia. The central-Pacific El Niño is characterized by positive SST anomalies in the central, rather than the eastern, equatorial Pacific.

**Walker circulation.** Walker circulation describes thermally direct, equatorial, zonal overturning circulation that converts available potential energy to kinetic energy of atmospheric motion. Over the Pacific Ocean, a zonal sea-level pressure gradient causes surface air to move from high pressure (caused by sinking motion) in the eastern Pacific to low pressure (caused by rising motion) in the western Pacific. Walker circulation is intrinsically connected to the ENSO, with a stronger flow during La Niña and a weaker flow during El Niño.

**Hadley circulation.** Hadley circulation describes tropical, meridional overturning circulation and is quantified by a mass-flux stream function. Strong diabatic heating near the thermal equator results in ascending air that spreads poleward, descends at subtropical high-pressure zones of both hemispheres, and flows towards the equator near the ocean or land surface. During the summer monsoon season, the Hadley circulation changes to a distinctly asymmetric flow, with ascent in the summer hemisphere and substantial cross-equatorial transport of energy and moisture<sup>28,98</sup>.

stalagmites that vary with 19-kyr or 23-kyr periodicity, similarly to precession, with an anti-phased interhemispheric relationship<sup>17,18</sup>. Recently, climatic interpretations of the stalagmite  $\delta^{18}\text{O}$  variability, particularly in the East Asian monsoon domain, have become controversial, with an overwhelming and growing number of observational and palaeoclimate data and model studies suggesting that the intensity of the local summer monsoon rainfall is not the only control on this proxy<sup>19–23</sup>.

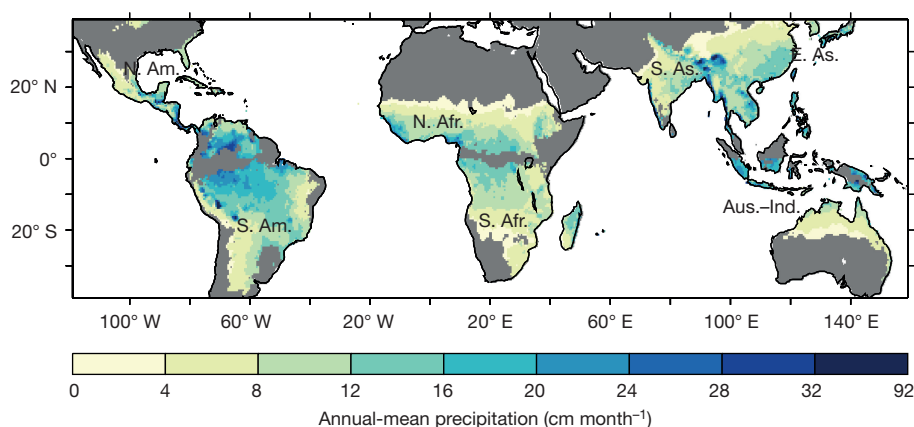
Despite this ongoing controversy surrounding the interpretation of stalagmite  $\delta^{18}\text{O}$  from the East Asian monsoon domain, model simulations

support the notion that precession is a strong control on rainfall in all monsoon domains (Fig. 3). During a precession minimum, the Northern Hemisphere summer is characterized by higher insolation and a stronger land–sea thermal gradient, which increase the atmospheric humidity (the thermodynamic component of monsoon rainfall) and wind circulation intensity (the dynamic component) (Box 1), whereas the Southern Hemisphere summer experiences the opposite scenario<sup>24</sup>. Simulations of changes in the spatial distribution of rainfall due to precession do not support a simple meridional shift in the seasonal position of the ITCZ, but do indicate a shift in the rainfall between land and ocean (Fig. 3a, b). Here, higher insolation causes the early-summer surface temperature to increase much faster over land than it does over the ocean, so that the maximum near-surface equivalent potential temperature (a quantity related to the stability of a column of air: if the equivalent potential temperature at the surface is greater than aloft, then the column is unstable and convection can occur) is shifted over land before the onset of the summer monsoon, resulting in enhanced land precipitation at the expense of rainfall over the adjacent sea<sup>20</sup>. By contrast, the maximum near-surface equivalent potential temperature and the precipitation centroid tend to stay over the ocean when insolation is low<sup>20</sup>.

Changes in obliquity with a period of 41 kyr affect the seasonality of incoming solar radiation equally in both hemispheres, with stronger variation in incoming solar radiation at high latitudes. Despite weak changes in incoming solar radiation at low latitudes, palaeoclimate data and model studies suggest that obliquity has a substantial effect on the strength of monsoon systems<sup>11,19,25,26</sup>, with increased summer monsoon rainfall when obliquity is maximal (Fig. 3c, d). A key factor in the traditional view of the (indirect) effect of obliquity on monsoons is high-latitude remote climate forcing associated with Northern Hemisphere ice sheets and sea-ice<sup>27</sup>. Obliquity-induced changes in cryosphere extent have been suggested to affect the monsoon domains by changing the oceanic and atmospheric circulations and the trajectories of moisture advection on glacial–interglacial timescales. Recent modelling results suggest that these changes are small for the South Asian monsoon, which becomes drier during glacial periods owing to lower temperatures<sup>23</sup> (a thermodynamic mechanism). However, for the East Asian monsoon domain, circulation changes are critically large and result in moisture convergence from the Pacific Ocean and increased precipitation, partly due to redistribution of air mass from the continents to the oceans as ice sheets grow, thereby enhancing the subtropical high-pressure system over the Pacific Ocean<sup>23</sup> (Fig. 4). In addition, the waxing and waning of ice affect the subtropical monsoons (Box 1) by changing the latitudinal temperature gradient and the position of the jet stream—a critical component of the East Asian monsoon<sup>22,28</sup>—and by inducing stationary planetary waves that strengthen the East Asian summer monsoon when ice grows<sup>29</sup>. The release of meltwater into the North Atlantic by waning ice sheets suppresses the North American monsoon more than any other monsoon domain<sup>30</sup>.

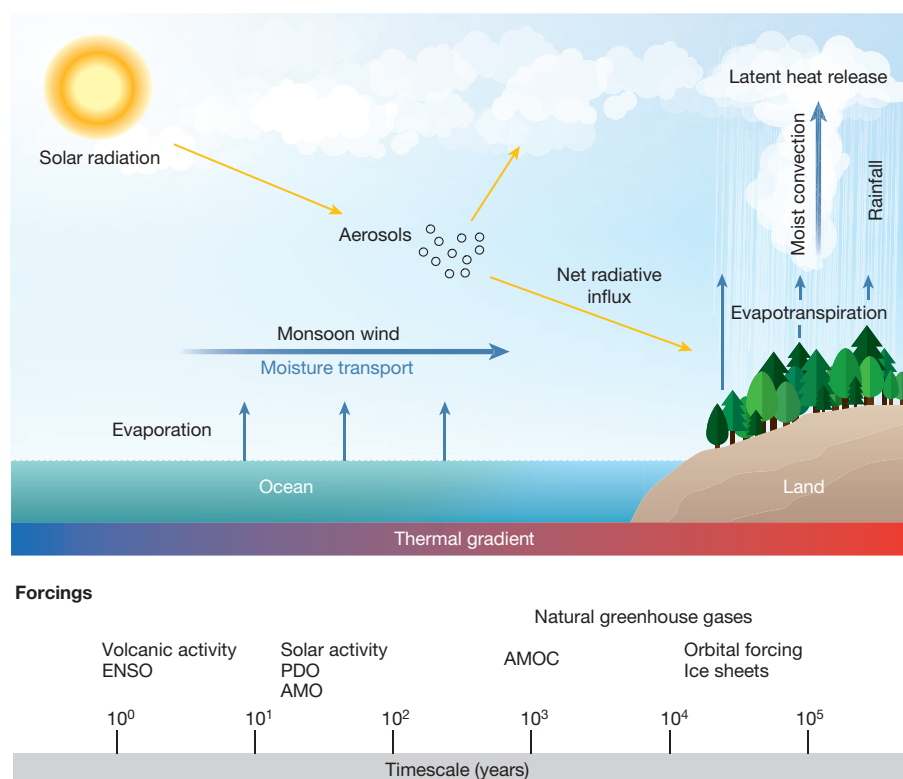
Recent studies suggest that obliquity also has a direct effect on monsoon rainfall by changing the meridional insolation gradient in the summer hemisphere and the interhemispheric insolation gradient<sup>26,31</sup>. The direct insolation-gradient forcing by obliquity can thus greatly contribute to changes in monsoon intensity at many locations without involving the (indirect) high-latitude remote climate forcing associated with Northern Hemisphere ice sheets<sup>25,26</sup> (Fig. 3). The presence of 41-kyr periodicity in West African monsoon records during the warm Pliocene<sup>27</sup>, that is, before the onset of strong Quaternary ice-sheet feedbacks in the Northern Hemisphere, supports the model-derived finding of a direct insolation-gradient forcing.

The timing of the response of different monsoon systems to precession and obliquity forcing and the degree to which they respond are matters of debate. Reconstructions of rainfall from the South American<sup>32</sup> and eastern North African<sup>11</sup> monsoons point to precession as the main forcing, whereas other reconstructions consider obliquity forcing to be at least of comparable importance for the West African<sup>11</sup> and South Asian<sup>11,19</sup> monsoons. On the other hand, model studies consistently suggest that, when ice sheet feedbacks are disabled, precession has a more severe



**Figure 1 | Global monsoon domain (coloured regions) as defined by the seasonality (summer–winter difference) in rainfall.** For the calculation of the monsoon domain, we applied the criteria of ref. 9—which require that the annual precipitation range (local summer–minus–winter precipitation) is larger than  $6 \text{ cm month}^{-1}$  and exceeds 50% of the annual-mean precipitation—except that the boreal summer was defined as June–August and the austral summer as December–February.

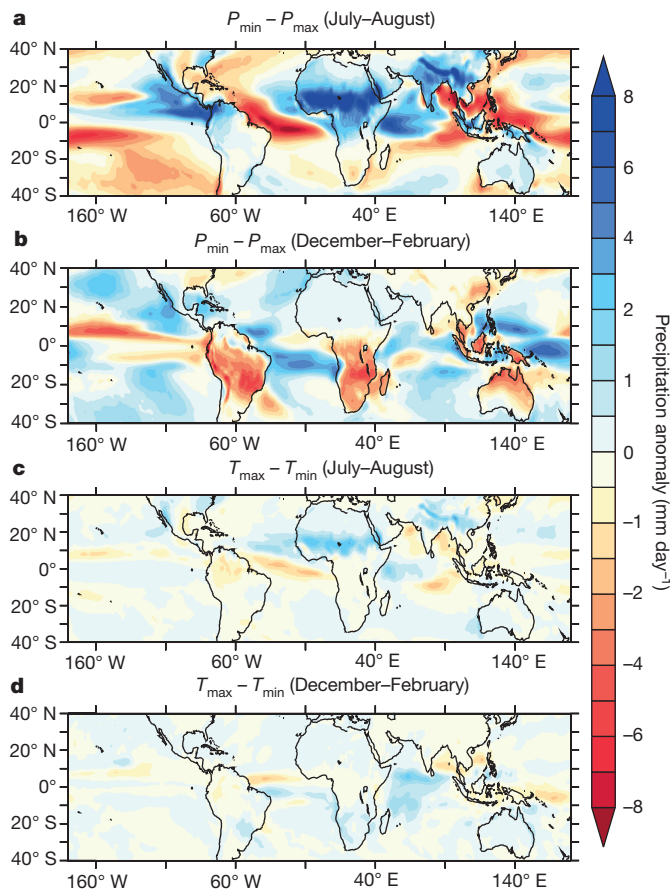
Colours denote annual-mean precipitation. Monthly global gridded, high-resolution ( $0.5^\circ$ ) precipitation data from 1901–2010 from the University of Delaware were used. Regional monsoon areas are: North America (N. Am.), South America (S. Am.), North Africa (N. Afr.), southern Africa (S. Afr.), South Asia (S. As.) or India, East Asia (E. As.) and Australia–Indonesia (Aus.–Ind.).



**Figure 2 | Basic components of a summer monsoon and its driving forces.** Summer solar radiation heats the surface. A land–ocean thermal contrast, resulting from a lower surface heat capacity of the land compared to the ocean, causes the low-level inflow of moist air from ocean to land during summer and the rising of air over the landmass. Recent studies of monsoon dynamics have de-emphasized the relevance of land–ocean surface contrasts and suggested that interactions between extratropical eddies and the tropical circulation are essential for the development of monsoons<sup>98</sup>. Condensation of water vapour in the rising air leads to the release of latent heat and rainfall. The latent-heat release in the troposphere reinforces the circulation and helps to pull in additional moisture from ocean to land. Depending on the vegetation cover, evapotranspiration on land may further feed the monsoon rain via local water recycling. Evapotranspiration may also affect the dynamics of the

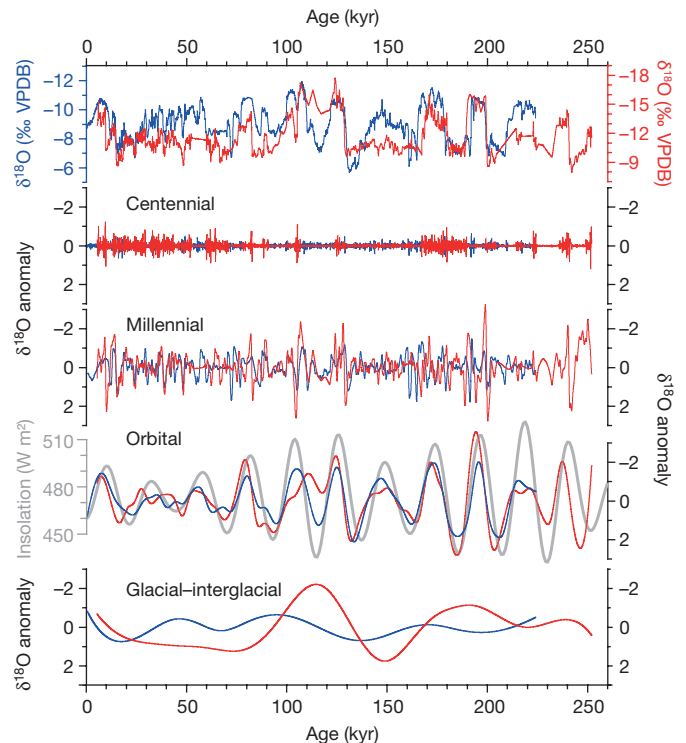
monsoon system via evaporative cooling of the land surface. Aerosols can affect the land–sea thermal contrast and, hence, the monsoon by modifying the transfer of solar radiation through the atmosphere by scattering and absorption processes. Aerosols may also have an effect on cloud microphysics, including changes in the radiative properties, frequency and lifetime of clouds. Because the various components of the monsoon system are closely coupled through feedback mechanisms, perturbation of any component may cause a chain reaction that affects the monsoon system as a whole. Different forcings perturb different components of the monsoon system, which then affect the entire system via feedback mechanisms, and act on different timescales. AMO, Atlantic Multidecadal Oscillation; AMOC, Atlantic meridional overturning circulation; ENSO, El Niño–Southern Oscillation; PDO, Pacific Decadal Oscillation.





**Figure 3 | Effects of obliquity and precession on tropical rainfall.** **a–d**, Precipitation anomalies (in millimetres per day; colour scale) as simulated by the high-resolution (1.125° horizontal resolution with 62 levels in the atmosphere), fully coupled, atmosphere–ocean general circulation model EC-Earth<sup>25</sup>.  $P_{\min} - P_{\max}$  (**a**, **b**) refers to the difference between the minimum and maximum climatic precession;  $T_{\max} - T_{\min}$  (**c**, **d**) refers to the difference between the maximum and minimum obliquity (axial tilt). Results are shown for boreal summer (June–August; **a**, **c**) and austral summer (December–February; **b**, **d**) mean precipitation. Applied minimum and maximum values for obliquity and climatic precession correspond to extreme values of the orbital parameters during the last one million years. Obliquity was set to a minimum value in **a** and **b**. In **c** and **d**, a circular orbit was assumed; that is, climatic precession was set to zero. During precession minimum ( $P_{\min}$ ), summer solstice occurs at perihelion such that seasonality is enhanced in the Northern Hemisphere and reduced in the Southern Hemisphere. During precession maximum ( $P_{\max}$ ), summer solstice occurs at aphelion. All other boundary conditions, such as the solar constant, greenhouse gas concentrations, sea level, ice sheets and vegetation, were kept fixed at pre-industrial levels during the model experiments. Details of the experimental set-up are described in ref. 25. Figure courtesy of J. H. C. Bosmans, University of Utrecht.

impact on monsoon rainfall than does obliquity<sup>25,33</sup> (Fig. 3), but without a consistent response of the specific monsoon systems to orbitally forced insolation changes. This is in part due to the varying influence of the internal feedback mechanisms such as ocean–atmosphere interaction. For instance, the sign and magnitude of sea surface temperature (SST) feedbacks on monsoon rainfall are equivocal in model simulations, with some implying a very limited effect on the South American<sup>34</sup> and East Asian<sup>35</sup> monsoons and others suggesting a large impact on the East Asian monsoon<sup>21</sup>. The influence of SST has been shown to amplify the insolation-induced strengthening of the African<sup>36</sup> and South Asian<sup>35</sup> monsoons, and to reduce the intensification of the South Asian monsoon<sup>36</sup>. Mechanistically, in some models, SST increases promote inland advection of moist air from the tropical Atlantic into the North African

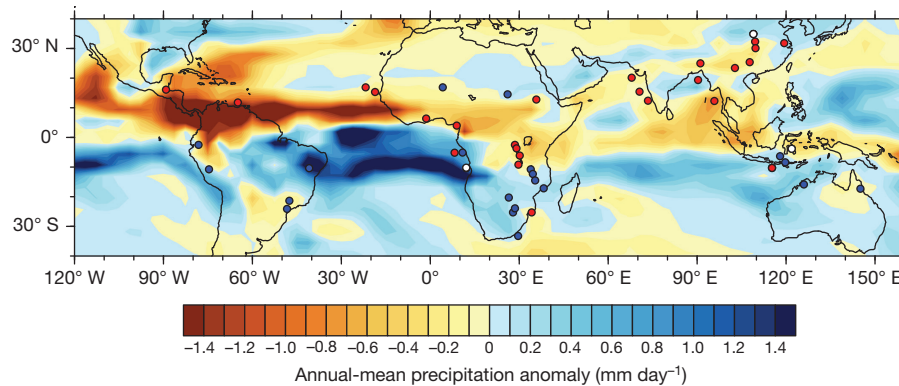


**Figure 4 | Monsoon variability at different timescales as evidenced by  $\delta^{18}\text{O}$  of cave stalagmites.** Ensemble empirical mode decomposition of cave speleothem  $\delta^{18}\text{O}$  records from the South Asian (red) and East Asian (blue) monsoon domains. Values are relative to Vienna Pee Dee Belemnite (VPDB). The bottom four panels show the intrinsic components decomposed from the speleothem  $\delta^{18}\text{O}$  records (top panel) ordered by timescales, as indicated, illustrating the stalagmite  $\delta^{18}\text{O}$  variance at each specific timescale. Summer (July–August) insolation at 30° N is shown in grey in the fourth panel. The lack of glacial–interglacial variability in the East Asian (blue) record in the bottom panel can be explained by the cancelling effect of increased precipitation (decrease in  $\delta^{18}\text{O}$ ) and a greater contribution of the Pacific moisture source (increase in  $\delta^{18}\text{O}$ ). Image adapted from ref. 23, National Academy of Sciences.

monsoon domain<sup>36</sup>, but shift the rainfall centroid towards the ocean in the South Asian monsoon domain<sup>20,36</sup>. By contrast, other models suggest that more divergent upper-tropospheric flow above the Arabian Sea and greater atmospheric water availability act as a positive feedback on the South Asian monsoon rainfall<sup>35</sup>.

The inconsistent and, in part, out-of-phase behaviour of the monsoon domains in response to insolation forcing is also due to their resonant response to insolation changes<sup>36</sup>. Transient simulations suggest that the subtropical monsoon systems are less sensitive to insolation forcing than are the tropical monsoons in terms of total rainfall, because the insolation-induced changes in summer rainfall are partly counter-balanced by precipitation changes in other seasons<sup>10</sup>. The sensitivity of the tropical monsoons is also not uniform: the North African monsoon is most sensitive to summer insolation, whereas the South Asian monsoon is most sensitive to spring-to-early-summer insolation<sup>33</sup>. It has been suggested that the reason for this non-uniformity is a resonant response of the South Asian monsoon to insolation forcing when maximum insolation anomalies occur near the summer solstice (for example, during the early Holocene) and a resonant response of the African monsoon—which has its rainfall maximum one month later in the annual cycle than the South Asian monsoon—when the maximum insolation change is delayed after the summer solstice (for example, during the middle Holocene)<sup>36</sup>.

In summary, it appears that the different responses of the individual monsoons to orbital forcing arise from their differing seasonal cycles that are influenced by regionally distinctive internal feedbacks. However, orbital forcing affects all the monsoon domains by changing



**Figure 5 | Monsoon rainfall anomalies during Heinrich stadial 1.**

Changes in rainfall during HS1 (15–17 kyr before present) relative to the Last Glacial Maximum (LGM, 19–21 kyr before present) as indicated by monsoon proxy records (filled circles) and simulated by the TraCE-21k climate model<sup>45</sup> (annual-mean precipitation anomalies, colour shading). The TraCE-21k simulation was performed with the coupled atmosphere–ocean general circulation model CCSM3 (Community Climate System Model, version 3). The resolution of the atmospheric component is 3.75°

in the horizontal direction with 26 layers in the vertical direction. Starting from the LGM, the model was integrated to the present-day, subject to realistically varying forcings by orbital insolation, atmospheric greenhouse gas concentrations, continental ice sheets, and meltwater fluxes. Filled red and blue circles indicate negative and positive monsoon rainfall anomalies, respectively. Filled white circles indicate no change or uncertain changes during HS1 compared to the LGM. Only records from monsoon domains as indicated in Fig. 1 are shown. For references, see main text.

the meridional gradient in insolation and, hence, heating; considering the recent differential hemispheric warming that will probably increase in the future<sup>4</sup>, exploring this aspect in palaeo-monsoon studies may help to better assess and predict future changes in monsoon rainfall. In addition, it has been suggested that obliquity-induced changes have some qualitative similarities with global-warming-induced changes, with respect to the summer-hemisphere Hadley cell driven by baroclinic eddies<sup>31</sup>. Therefore, a deeper understanding of obliquity-induced changes in tropical circulation is crucial for more reliable predictions of climate change.

### North Atlantic forcing

Today, a northward cross-equatorial, ocean heat transport of about 0.5 PW, which is mainly due to the Atlantic meridional overturning circulation (AMOC), results in relatively warm surface waters in the northern North Atlantic and renders the Northern Hemisphere warmer than the Southern Hemisphere. In response, there is a cross-equatorial, southward transport of energy via the mean Hadley circulation, as a result of the ascending branch of the Hadley circulation and the mean position of its ITCZ being located north of the Equator<sup>37</sup>. Any physical mechanism that induces anomalous warming or cooling of one hemisphere relative to the other will cause a shift in the mean position of the ITCZ towards the warming hemisphere<sup>38,39</sup>, which will affect tropical monsoon systems.

Heinrich and Dansgaard–Oeschger events of the last glacial period serve as paradigms of oscillations in interhemispheric temperature asymmetry<sup>40</sup>. During these millennial-scale climate fluctuations, changes in the interhemispheric thermal gradient were probably caused by variations in the AMOC<sup>41</sup> and its associated heat transport, and amplified by sea-ice feedbacks<sup>42</sup>. As a result, the largest temperature anomalies appeared in the North Atlantic realm<sup>40,42</sup>; nevertheless, millennial-scale glacial climate variability is well expressed in numerous palaeo-monsoon records from all over the globe<sup>17,18,43</sup> (Fig. 4).

The largest body of data exists for the last Heinrich stadial (HS1, about 18–15 kyr before present) when the AMOC was nearly shut off and surface temperatures in the North Atlantic realm were at a minimum<sup>44</sup>. A compilation of available palaeo-hydrological records for HS1 suggests a general pattern of a global-scale, southward ITCZ shift associated with changing interhemispheric thermal asymmetry (Fig. 5). Specifically, substantial drying in the North African and South Asian monsoon regions is found, whereas the Australian–Indonesian and South American monsoon systems become wetter. Speleothem isotope records from the East Asian monsoon system also indicate dry conditions during HS1, although model studies cast doubt on the local-rainfall interpretation of these records (see above). The pattern of HS1 precipitation response

for the southern African monsoon region appears to be more complex. While the northern portion of the southern African monsoon region becomes drier, the southern portion witnesses anomalously wet conditions during HS1 (Fig. 5).

Owing to the global distribution of HS1 monsoon records, model–data comparison for this climate event offers a unique possibility for testing the reliability of climate model simulations with respect to AMOC slowdown. The TraCE-21k transient simulation of the last deglaciation supports the notion that AMOC intensity is a dominant control on tropical rainfall patterns on millennial timescales<sup>45,46</sup>. A comparison of the HS1 hydroclimatic records with the TraCE-21k-simulated HS1 precipitation response shows agreement in terms of the sign of the rainfall anomalies over most monsoon regions (Fig. 5). However, discrepancies between the model and the data stand out in southeastern Brazil, close to the South Atlantic convergence zone. Furthermore, reconstructed dry conditions in the South Asian monsoon region appear to be greatly underestimated by TraCE-21k. The reason for this weak rainfall response in the model is still uncertain, but we are beginning to understand the teleconnections between North Atlantic SST forcing and the South Asian summer monsoon. Several studies have examined possible teleconnections, and different mechanisms have been proposed, including a tropospheric temperature anomaly over Eurasia related to the North Atlantic Oscillation<sup>47</sup>, a pathway over the Pacific Walker circulation<sup>48</sup>, a southward shift of the subtropical westerly jet over Africa and Asia<sup>49</sup>, and a wave-like atmospheric teleconnection between the northern North Atlantic and India<sup>50</sup>.

As well as abrupt climate events during the last glacial period and deglaciation, several palaeo-monsoon records also witness centennial-to-millennial-scale climate variability during the Holocene associated with North Atlantic cold phases, such as the effect of the North Atlantic cold event that took place 8.2 kyr ago and is akin to the HS1 rainfall anomaly pattern<sup>17,18,43</sup>. Operating on a much shorter timescale, the Atlantic Multidecadal Oscillation (AMO; Box 1) is a mode of natural climate variability characterized by a coherent pattern of variability in basin-wide North Atlantic SST with a period of 60–80 years. Observational data suggest that North Atlantic SST associated with the AMO has an influence on multidecadal global monsoon variability<sup>51</sup>, with the overall pattern of changes in global monsoon rainfall during the cold phase of the AMO resembling the HS1 anomaly pattern (Fig. 5). However, owing to the shortness of the instrumental record, the robustness of these links is difficult to assess. High-resolution proxy records enable us to extend the instrumental monsoon records to the past few thousand years<sup>30,52–54</sup>.



In view of a potential slowdown of the AMOC in the coming decades<sup>55</sup>, understanding the connection between oceanic heat transport and monsoon systems is of paramount importance for predicting the fate of the monsoons. Meanwhile, a growing body of palaeo-monsoon records allows us to paint a global picture of changes in monsoon rainfall in response to large-scale oceanic redistribution of heat.

## CO<sub>2</sub> forcing

Models predict an increase in total monsoon rainfall during the twenty-first century in response to rising atmospheric greenhouse gas (GHG) forcing, increasing atmospheric moisture content, and an expansion of the area affected by monsoons<sup>1</sup>. Isolating the effects of atmospheric GHG forcing on monsoon dynamics from palaeoclimatic records is difficult because changes in GHG concentrations and other forcings usually occurred contemporaneously. Not only did GHG concentrations oscillate in concert with Quaternary glacial–interglacial cycles, but so did the size of ice sheets, the area of sea-ice, and the orbital forcing. Traditionally, dry conditions in North Africa during Quaternary glacial stages have been attributed to expansion of the boreal cryosphere and North Atlantic cooling<sup>27</sup>. However, recently published results from the TraCE-21k transient simulation of the last glacial termination suggest that the GHG forcing plays an important role in developing wet interglacial conditions in the African monsoon regions<sup>46</sup>. Conversely, the lack of glacial–interglacial variability in rainfall proxy records from deep tropical monsoon systems that reach beyond the last glacial termination indicates that GHG forcing plays a limited role and that the boreal cryosphere has a stronger influence<sup>56</sup>.

The Pliocene warm period with a relatively high atmospheric CO<sub>2</sub> concentration is often considered as a potential analogue of future climate. Palaeoclimatic evidence suggests wet Pliocene conditions in the West African monsoon region, including the presence of Saharan palaeo-rivers, while forests, woodland and savannah extended further north compared to the Pleistocene<sup>57</sup>. For the East Asian monsoon region, proxy data and climate models suggest stronger-than-modern summer winds associated with wetter conditions during the mid-Pliocene warm period<sup>58</sup>. To what extent the wetter Pliocene conditions can be attributed to direct GHG forcing of the monsoons remains uncertain, because other potential forcing factors (for example, ice sheets, topography and global ocean circulation) were also different from today. Yet additional Pliocene records of rainfall from other monsoon regions will help identifying the sign of change in a warmer-than-today Earth. Isolating the effects of atmospheric GHG forcing on monsoon dynamics remains a critical task for palaeo-monsoon studies, to narrow down uncertainties in projections of future monsoon rainfall.

## ENSO and PDO

Similarly to the AMO (see above), instrumental records suggest that SST anomalies in the tropical Pacific associated with the ENSO are a predominant forcing of monsoon variability in modern climate<sup>3</sup>. ENSO-related SST anomalies affect the global atmospheric circulation, particularly the Walker circulation (Box 1). Generally, the east–west displacement of the ascending and descending branches of the Walker circulation during El Niño years results in an increased descent over Australasia and reduced monsoon rainfall<sup>59</sup>. However, this relationship appears unstable in the instrumental records, partly as a result of the different types of the ENSO, which are characterized by spatially varying SST anomalies (between the central and eastern equatorial Pacific; Box 1) and teleconnections<sup>60</sup>, and highlights the need for palaeoclimatic evidence beyond the instrumental records.

Proxy records of monsoon rainfall in North America support a leading role of the ENSO and the Pacific Decadal Oscillation (PDO; Box 1) after the Northern Hemisphere summer and autumn insolation declined about 4,000 years ago by changing the Pacific SST<sup>61</sup>, the Pacific–Atlantic SST gradient<sup>62</sup>, or the position of the ITCZ<sup>63</sup>. Whether the observed changes in monsoon precipitation are caused by a combination of forcings<sup>30</sup>, the ENSO only<sup>61,64</sup>, or solar forcing<sup>65</sup> remains controversial. This is in part

due to the fact that different forcings give rise to similar responses within the same monsoon domain. For example, La Niña, positive AMO and negative PDO cause a wet Mesoamerica and dry southwest USA, whereas El Niño, negative AMO and positive PDO give rise to a dry Mesoamerica and wet southwest USA. More importantly, the suggested links to ENSO, PDO or AMO rely mainly on simple statistical analyses of the rainfall proxy data such as power spectra or correlation coefficients that are not sufficient to infer any causal mechanisms.

Palaeoclimate data from the Australasian monsoon regions imply that the ENSO has a heterogeneous spatiotemporal impact even within the same monsoon domain. A 50-year stalagmite record from northeast India suggests that El Niño and positive PDO have diminished monsoon precipitation in central India and shortened the moisture transport pathways to northeast India during the past decades<sup>66</sup>. Tree-ring data of the last millennium suggest that a simple canonical form of ENSO influence is insufficient to explain Asian monsoon variability, probably owing to the different pathways through which ENSO interacts with the different components of the monsoon<sup>67</sup>. Secular changes in ENSO teleconnections further complicate a general assessment of ENSO-related rainfall variability and make it somewhat regionally dependent<sup>1</sup>. Finally, the lack of indisputable and continuous benchmark records for the ENSO, PDO and AMO is another obstacle in untangling their relationship to monsoons. While many independent reconstructions of the AMOC, atmospheric CO<sub>2</sub> concentration and solar activity are available, data relating to the variability in the ENSO, PDO and AMO, in forcings and responses, and in lead and lag times on scales longer than a few decades do not exist. Therefore, reconstructing past changes in the monsoon–ENSO–PDO relationship remains one of the most critical contributions of palaeoclimate research to the climate change discussion.

## Land cover

Land-cover changes alter surface roughness, albedo and water fluxes, thereby affecting the energy and moisture budgets of monsoon systems (Fig. 2). Over the Tibetan plateau, expansion of vegetation has been observed in response to recent warming, leading to enhanced evaporative cooling with potential effects on the South Asian monsoon<sup>68</sup>. On the other hand, deforestation in Mexico has been identified as a drought amplifier<sup>69</sup>, and human-induced land-cover changes in China over the past 3,400 years have been suggested to weaken the East Asian summer monsoon<sup>70</sup>. Observational evidence from the period 1981–2003 suggests that the decrease in the South Asian monsoon rainfall may have been caused by agricultural intensification<sup>71</sup>, and model simulations indicate that land-use changes and deforestation may cause a locally delayed<sup>72</sup> and reduced monsoon rainfall in the Northern Hemisphere by shifting the ITCZ southward<sup>73</sup>.

Although anthropogenic land-cover changes may act as a forcing for shifts in monsoons, a dynamic vegetation cover may also act as a feedback to changes in monsoon rainfall, especially in North Africa. Pioneering work by Charney<sup>74</sup> has suggested a strong positive feedback between vegetation cover and monsoon rainfall in the Sahel region. On the basis of these ideas it has been proposed that the effect of expanded North African vegetation cover on surface albedo would have been crucial in amplifying the orbitally triggered, early-to-mid Holocene, West African monsoon rainfall anomaly—the so-called African humid period, during which the Sahara was much ‘greener’ than it is today<sup>15</sup>. Provided that the positive vegetation–rainfall feedback is strong enough to introduce nonlinear dynamics into the climate–vegetation system, two equilibria of the regional atmosphere–vegetation state may exist: a humid/green state and a dry/desert state. A transition from the humid state to the dry state by a catastrophic bifurcation was suggested to have abruptly terminated the African humid period around 5.5 kyr ago<sup>16</sup>. Other studies have questioned both the abruptness of the large-scale North African climate transition<sup>75,76</sup> and the existence of a strong local Charney feedback, which operates through changes in surface albedo<sup>77</sup>. Instead, a recent model study has suggested a positive vegetation–rainfall feedback

during the Holocene that operated via anomalies in surface latent-heat flux caused by canopy evaporation and transpiration and their effect on the mid-tropospheric African easterly jet<sup>78</sup>. Another study highlighted the role of remote forcing from expanded forest cover in Eurasia in amplifying North African rainfall during the early-to-mid Holocene<sup>79</sup>. Both feedback mechanisms are specific to the North African monsoon and cannot be generalized to the other monsoon regions<sup>35</sup>.

In summary, forcing and feedbacks associated with land-cover changes may have important effects on the different regional monsoon systems. Even though quantification of monsoon–vegetation feedbacks by means of proxy data alone will be difficult, reconstructions of past land and vegetation cover are indispensable to test model-derived hypotheses and may aid in the improvement of parameterizations in land surface models.

### Solar and volcanic forcing

Solar and volcanic activities are considered major external forcings of climate variability for the last millennium and beyond<sup>80,81</sup>. Several studies suggest that short-term changes in solar activity, such as the 11-year sunspot cycle, may affect monsoon intensity. In some parts of the South Asian monsoon domain, the average precipitation anomalies in the five most-recent peaks in the sunspot cycle reached values as high as 20% above normal, and suggest that solar activity influences monsoon strength<sup>82</sup>. A response to the 11-year solar cycle was also found for the East Asian monsoon<sup>83</sup>, with the band of heaviest rain penetrating further north when sunspot numbers were high. However, observations of sun–monsoon relationships are based on relatively short time series and, hence, are inherently uncertain. High-resolution speleothem records from Northern Hemisphere monsoon domains suggest a positive correlation between solar activity and Holocene monsoon rainfall on longer timescales<sup>65,84,85</sup>. Because changes in solar forcing are rather small, any noticeable effect on climate requires nonlinear responses and amplifying feedbacks<sup>80</sup>. ‘Top-down’ processes, based on solar heating of the stratosphere and changes in stratospheric ozone concentration via modification of photo-dissociation rates, are considered to be crucial to most climatic phenomena associated with solar forcing<sup>80</sup>. Yet ‘bottom-up’ processes, acting via solar-induced changes in SST, may also be important in the tropics, owing to positive feedbacks associated with surface evaporation, moisture fluxes and latent-heat release in monsoon systems<sup>86,87</sup>.

Recent observations, palaeoclimate data and model studies show that volcanic forcing is probably more important than solar forcing on a hemispheric-to-global scale, and drove a large portion of the interannual-to-multidecadal monsoon variability during the late Holocene by affecting the SST, ENSO and AMO<sup>81,88–90</sup>. The primary effect of volcanic sulfate aerosols in the stratosphere is to cool the surface of Earth by reflecting incoming solar radiation. As post-eruption cooling is generally stronger over land than it is over the ocean, a weaker summer monsoon circulation should result from a reduced land–ocean thermal contrast (see Fig. 2). Model results suggest that, after large volcanic eruptions, cooling over east Asia is stronger than that over the tropical ocean, favouring weaker East Asian summer monsoon circulation<sup>91</sup>. However, tree-ring data of the past 750 years contradict model results regarding the sign of the monsoon response in the year of eruption, and suggest an anomalously wet southeast Asia and dry conditions over central Asia<sup>92</sup>. On a global scale, large, explosive volcanic eruptions may cause reorganizations of the Hadley cell and a shift of the ITCZ away from the hemisphere with the greater aerosol concentration<sup>90</sup>.

The magnitudes of solar and volcanic forcing and their effects on the hydrologic cycle are sources of large uncertainty in most model simulations, owing to the common model deficiencies in capturing precipitation variability at decadal-to-multi-decadal timescales as well as chemical and physical processes related to aerosol forcing<sup>89</sup>; overcoming these deficiencies in the next generation of climate models is critical. A growing body of evidence from palaeoclimate data underpins a strong influence of solar and volcanic forcing on monsoon variability, and may help to better identify these model deficiencies.

### Outlook

Various forcings acting on different timescales affect the monsoon systems on a global scale (Figs 2, 4); regional monsoon responses differ from global responses in terms of sensitivity and timing, which can be attributed to different feedbacks and teleconnections. Palaeoclimatic records may help to elucidate the regional characteristics of the different monsoon subsystems and to better understand the internal feedback processes within the climate system; these feedback processes are suggested to influence the recent monsoon circulation more severely than the external forcing by increasing GHGs<sup>3</sup>. In addition, palaeoclimatic records provide valuable information that can be used to assess the ability of climate models to simulate monsoon changes for different climates, and are of utmost importance in testing model-derived hypotheses. For instance, a recent model study highlighted the role of differential warming between the Northern Hemisphere extratropics and tropics in determining the future development of the North African monsoon, where strong extratropical warming will induce a substantial increase in Sahel rainfall<sup>93</sup>. A large obliquity signature in the North African palaeo-monsoon records<sup>27</sup> strongly supports this model-derived finding, because obliquity forcing involves a substantial change in the meridional gradients of insolation and heating in the summer hemisphere<sup>31</sup> as well as amplification of extratropical temperature change via cryosphere expansion and retraction<sup>27</sup>. In a more general sense, it has been suggested that obliquity-induced changes have some qualitative similarities with global-warming-induced changes, with respect to the summer-hemisphere Hadley cell and ITCZ shifts related to mid-latitude eddy activity<sup>31</sup>. Therefore, a deeper understanding of obliquity-induced changes in tropical and extratropical circulation may help to better assess and predict future changes in monsoon rainfall.

Climate models predict that the ongoing interhemispheric thermal asymmetry will lead to a northward shift of the mean ITCZ position<sup>4</sup>, but accurate projection of monsoon rainfall suffers from uncertainties in SST warming<sup>13</sup>. Because the location of the ITCZ is sensitive to changes in the meridional SST gradient<sup>12</sup>, which can be constrained using palaeo-thermometric methods, palaeoclimate reconstructions that involve changes in (inter)hemispheric SST gradients provide unambiguous evidence for such ITCZ migrations and a blueprint of the associated changes in monsoon rainfall patterns.

Changes in monsoon rainfall can be thermodynamic or dynamic<sup>24</sup>, and may be related to changes in the length of the rainy season, the rainfall intensity or the monsoon area. The various aspects of changes in monsoon rainfall require a deeper understanding of the mechanisms behind palaeo-monsoon variability, which may be achieved by means of palaeoclimatic monsoon reconstructions that use multiple proxy records, ideally from the same location<sup>94</sup>. Combining proxies for wind and precipitation changes may help to estimate the relative importance of the dynamic and thermodynamic aspects of changes in monsoon rainfall. A promising approach for distinguishing past changes in wet-season intensity from changes in wet-season length has been proposed, which involves combining leaf-wax  $\delta D$  with leaf-wax  $\delta^{13}C$  analyses<sup>95</sup>.

Palaeoclimatic data might contain further quantitative information on the magnitude of change that could help to better project future climate variations. Climate model experiments suggest that the observed decrease in South Asian monsoon rainfall over the last 70 years is mostly attributable to anthropogenic aerosol emissions<sup>96</sup>. Water-isotope-enabled climate models will be essential for the appropriate interpretation of water-isotope records and a quantitative reconstruction of past monsoon rainfall variability; however, reducing the uncertainty related to atmospheric aerosol loading and to the associated teleconnections, including the ENSO, PDO and AMO, is of equal importance for achieving more trustworthy climate projections. In particular, it has been shown that eastern and central Pacific ENSO events have different effects on the South Asian<sup>59</sup> and East Asian<sup>97</sup> monsoons. Indisputable and continuous records of the different states (El Niño and La Niña) and types (central and eastern Pacific) of the ENSO and of the PDO and AMO are critical to evaluate the stability of monsoon teleconnections. Finally, the middle Pliocene is probably



the best available analogue for future warming and, thus, palaeoclimate records of the warm, high-CO<sub>2</sub> Earth of the Pliocene are urgently needed to narrow down uncertainties in future monsoon projections.

Received 30 September 2015; accepted 24 February 2016.

1. Christensen, J. H. *et al.* in *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change* (eds Stocker, T. F. *et al.*) Ch. 14, 1217–1308 (Cambridge Univ. Press, 2013).
2. Flato, G. *et al.* in *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change* (eds Stocker, T. F. *et al.*) Ch. 9, 741–866 (Cambridge Univ. Press, 2013).
3. Wang, B. *et al.* Northern Hemisphere summer monsoon intensified by mega-El Niño/southern oscillation and Atlantic multidecadal oscillation. *Proc. Natl Acad. Sci. USA* **110**, 5347–5352 (2013).
4. Lee, J.-Y. & Wang, B. Future change of global monsoon in the CMIP5. *Clim. Dyn.* **42**, 101–119 (2014).
5. Sperber, K. R. *et al.* The Asian summer monsoon: an intercomparison of CMIP5 vs. CMIP3 simulations of the late 20th century. *Clim. Dyn.* **41**, 2711–2744 (2013).
6. Wang, B. *et al.* Rethinking Indian monsoon rainfall prediction in the context of recent global warming. *Nature Commun.* **6**, 7154 (2015).
7. Asharaf, S. & Ahrens, B. Indian summer monsoon rainfall processes in climate change scenarios. *J. Clim.* **28**, 5414–5429 (2015).
8. Trenberth, K. E., Stepaniak, D. P. & Caron, J. M. The global monsoon as seen through the divergent atmospheric circulation. *J. Clim.* **13**, 3969–3993 (2000).
9. Wang, B. & Ding, Q. Global monsoon: dominant mode of annual variation in the tropics. *Dyn. Atmos. Oceans* **44**, 165–183 (2008).
10. Dallmeyer, A. *et al.* The evolution of sub-monsoon systems in the Afro-Asian monsoon region during the Holocene—comparison of different transient climate model simulations. *Clim. Past* **11**, 305–326 (2015).
11. Caley, T. *et al.* Orbital timing of the Indian, East Asian and African boreal monsoons and the concept of a ‘global monsoon’. *Quat. Sci. Rev.* **30**, 3705–3715 (2011).
12. Donohoe, A., Marshall, J., Ferreira, D. & McGee, D. The relationship between ITCZ location and cross-equatorial atmospheric heat transport: from the seasonal cycle to the Last Glacial Maximum. *J. Clim.* **26**, 3597–3618 (2013).
13. Chen, X. & Zhou, T. Distinct effects of global mean warming and regional sea surface warming pattern on projected uncertainty in the South Asian summer monsoon. *Geophys. Res. Lett.* **42**, 9433–9439 (2015).
14. Kutzbach, J. E. Monsoon climate of the early Holocene: climate experiment with the Earth’s orbital parameters for 9000 years ago. *Science* **214**, 59–61 (1981); erratum 214, 606 (1981).
15. Claussen, M. & Gayler, V. The greening of the Sahara during the mid-Holocene: results of an interactive atmosphere-biome model. *Global Ecol. Biogeogr. Lett.* **6**, 369–377 (1997).
16. deMenocal, P. *et al.* Abrupt onset and termination of the African Humid Period: rapid climate responses to gradual insolation forcing. *Quat. Sci. Rev.* **19**, 347–361 (2000).
17. Cheng, H., Sinha, A., Wang, X., Cruz, F. & Edwards, R. The Global Paleomonsoon as seen through speleothem records from Asia and the Americas. *Clim. Dyn.* **39**, 1045–1062 (2012).
18. Wang, P. X. *et al.* The global monsoon across timescales: coherent variability of regional monsoons. *Clim. Past* **10**, 2007–2052 (2014).
19. Clemens, S. C., Prell, W. L. & Sun, Y. Orbital-scale timing and mechanisms driving late Pleistocene Indo-Asian summer monsoons: reinterpreting cave speleothem  $\delta^{18}\text{O}$ . *Paleoceanography* **25**, PA4207 (2010).
20. Battisti, D. S., Ding, Q. & Roe, G. H. Coherent pan-Asian climatic and isotopic response to orbital forcing of tropical insolation. *J. Geophys. Res. Atmospheres* **119**, 11997–12020 (2014).
21. Caley, T., Roche, D. M. & Renssen, H. Orbital Asian summer monsoon dynamics revealed using an isotope-enabled global climate model. *Nature Commun.* **5**, 5371 (2014).
22. Chiang, J. C. H. *et al.* Role of seasonal transitions and westerly jets in East Asian paleoclimate. *Quat. Sci. Rev.* **108**, 111–129 (2015).
23. Cai, Y. *et al.* Variability of stalagmite-inferred Indian monsoon precipitation over the past 252,000 y. *Proc. Natl Acad. Sci. USA* **112**, 2954–2959 (2015).
24. Merlis, T. M., Schneider, T., Bordoni, S. & Eisenman, I. The tropical precipitation response to orbital precession. *J. Clim.* **26**, 2010–2021 (2013).
25. Bosmans, J. H. C., Drijfhout, S. S., Tuenter, E., Hilgen, F. J. & Lourens, L. J. Response of the North African summer monsoon to precession and obliquity forcings in the EC-Earth GCM. *Clim. Dyn.* **44**, 279–297 (2015).
26. Bosmans, J. H. C., Hilgen, F. J., Tuenter, E. & Lourens, L. J. Obliquity forcing of low-latitude climate. *Clim. Past* **11**, 1335–1346 (2015).
27. deMenocal, P. B. Plio-Pleistocene African climate. *Science* **270**, 53–59 (1995).
28. Molnar, P., Boos, W. R. & Battisti, D. S. Orographic controls on climate and paleoclimate of Asia: thermal and mechanical roles for the Tibetan plateau. *Annu. Rev. Earth Planet. Sci.* **38**, 77–102 (2010).
29. Muri, H., Berger, A., Yin, Q., Karami, M. P. & Barriat, P.-Y. The climate of the MIS-13 interglacial according to HadCM3. *J. Clim.* **26**, 9696–9712 (2013).
30. Metcalfe, S. E., Barron, J. A. & Davies, S. J. The Holocene history of the North American Monsoon: ‘known knowns’ and ‘known unknowns’ in understanding its spatial and temporal complexity. *Quat. Sci. Rev.* **120**, 1–27 (2015).
31. Mantsis, D. F. *et al.* The response of large-scale circulation to obliquity-induced changes in meridional heating gradients. *J. Clim.* **27**, 5504–5516 (2014).
32. Baker, P. A. & Fritz, S. C. Nature and causes of Quaternary climate variation of tropical South America. *Quat. Sci. Rev.* **124**, 31–47 (2015).
33. Rachmayani, R., Prange, M. & Schulz, M. Intra-interglacial climate variability: model simulations of Marine Isotope Stages 1, 5, 11, 13, and 15. *Clim. Past* **12**, 677–695 (2016).
34. Liu, X. & Battisti, D. S. The influence of orbital forcing of tropical insolation on the climate and isotopic composition of precipitation in South America. *J. Clim.* **28**, 4841–4862 (2015).
35. Dallmeyer, A., Claussen, M. & Otto, J. Contribution of oceanic and vegetation feedbacks to Holocene climate change in monsoonal Asia. *Clim. Past* **6**, 195–218 (2010).
36. Braconnot, P., Marzin, C., Grégoire, L., Mosquet, E. & Marti, O. Monsoon response to changes in Earth’s orbital parameters: comparisons between simulations of the Eemian and of the Holocene. *Clim. Past* **4**, 281–294 (2008).
37. Schneider, T., Bischoff, T. & Haug, G. H. Migrations and dynamics of the intertropical convergence zone. *Nature* **513**, 45–53 (2014).
38. Chiang, J. H. & Bitz, C. Influence of high latitude ice cover on the marine Intertropical Convergence Zone. *Clim. Dyn.* **25**, 477–496 (2005).
39. Chiang, J. C. H., Biasutti, M. & Battisti, D. S. Sensitivity of the Atlantic Intertropical Convergence Zone to Last Glacial Maximum boundary conditions. *Paleoceanography* **18**, 1094 (2003).
40. Rahmstorf, S. Ocean circulation and climate during the past 120,000 years. *Nature* **419**, 207–214 (2002).
41. Gottschalk, J. *et al.* Abrupt changes in the southern extent of North Atlantic Deep Water during Dansgaard-Oeschger events. *Nature Geosci.* **8**, 950–954 (2015).
42. Zhang, X., Prange, M., Merkel, U. & Schulz, M. Spatial fingerprint and magnitude of changes in the Atlantic meridional overturning circulation during marine isotope stage 3. *Geophys. Res. Lett.* **42**, 1903–1911 (2015).
43. Zhiseng, A. *et al.* Global monsoon dynamics and climate change. *Annu. Rev. Earth Planet. Sci.* **43**, 29–77 (2015).
44. McManus, J. F., Francois, R., Gherardi, J. M., Keigwin, L. D. & Brown-Leger, S. Collapse and rapid resumption of Atlantic meridional circulation linked to deglacial climate changes. *Nature* **428**, 834–837 (2004).
45. Liu, Z. *et al.* Transient simulation of last deglaciation with a new mechanism for Bølling-Allerød warming. *Science* **325**, 310–314 (2009).
46. Otto-Bliessner, B. L. *et al.* Coherent changes of southeastern equatorial and northern African rainfall during the last deglaciation. *Science* **346**, 1223–1227 (2014).
47. Goswami, B. N., Madhusoodanan, M. S., Neema, C. P. & Sengupta, D. A physical mechanism for North Atlantic SST influence on the Indian summer monsoon. *Geophys. Res. Lett.* **33**, L02706 (2006).
48. Zhang, R. & Delworth, T. L. Simulated tropical response to a substantial weakening of the Atlantic thermohaline circulation. *J. Clim.* **18**, 1853–1860 (2005).
49. Marzin, C., Kallel, N., Kageyama, M., Duplessy, J.-C. & Braconnot, P. Glacial fluctuations of the Indian monsoon and their relationship with North Atlantic climate: new data and modelling experiments. *Clim. Past* **9**, 2135–2151 (2013).
50. Mohtadi, M. *et al.* North Atlantic forcing of tropical Indian Ocean climate. *Nature* **509**, 76–80 (2014).
51. Ting, M., Kushnir, Y., Seager, R. & Li, C. Robust features of Atlantic multi-decadal variability and its climate impacts. *Geophys. Res. Lett.* **38**, L17705 (2011).
52. Berkelhammer, M. *et al.* Persistent multidecadal power of the Indian Summer Monsoon. *Earth Planet. Sci. Lett.* **290**, 166–172 (2010).
53. Vuille, M. *et al.* A review of the South American monsoon history as recorded in stable isotopic proxies over the past two millennia. *Clim. Past* **8**, 1309–1321 (2012).
54. Shanahan, T. M. *et al.* Atlantic forcing of persistent drought in West Africa. *Science* **324**, 377–380 (2009).
55. Rahmstorf, S. *et al.* Exceptional twentieth-century slowdown in Atlantic Ocean overturning circulation. *Nature Clim. Change* **5**, 475–480 (2015).

56. Russell, J. M. *et al.* Glacial forcing of central Indonesian hydroclimate since 60,000 y B.P. *Proc. Natl Acad. Sci. USA* **111**, 5100–5105 (2014).
57. Vallé, F., Dupont, L. M., Leroy, S. A. G., Schefuß, E. & Wefer, G. Pliocene environmental change in West Africa and the onset of strong NE trade winds (ODP Sites 659 and 658). *Palaeogeogr. Palaeoclimatol. Palaeoecol.* **414**, 403–414 (2014).
58. Zhang, R. *et al.* Mid-Pliocene East Asian monsoon climate simulated in the PlioMIP. *Clim. Past* **9**, 2085–2099 (2013).
59. Kumar, K. K., Rajagopalan, B., Hoerling, M., Bates, G. & Cane, M. Unraveling the mystery of Indian monsoon failure during El Niño. *Science* **314**, 115–119 (2006).
60. Yeh, S.-W. *et al.* El Niño in a changing climate. *Nature* **461**, 511–514 (2009).
61. Lachniet, M. S., Bernal, J. P., Asmerom, Y., Polyak, V. & Piperno, D. A 2400 yr Mesoamerican rainfall reconstruction links climate and cultural change. *Geology* **40**, 259–262 (2012).
62. Douglas, P. M. J. *et al.* Drought, agricultural adaptation, and sociopolitical collapse in the Maya Lowlands. *Proc. Natl Acad. Sci. USA* **112**, 5607–5612 (2015).
63. Bernal, J. P. *et al.* A speleothem record of Holocene climate variability from southwestern Mexico. *Quat. Res.* **75**, 104–113 (2011).
64. Polissar, P. J., Abbott, M. B., Wolfe, A. P., Vuille, M. & Bezada, M. Synchronous interhemispheric Holocene climate trends in the tropical Andes. *Proc. Natl Acad. Sci. USA* **110**, 14551–14556 (2013).
65. Asmerom, Y., Polyak, V. J., Rasmussen, J. B. T., Burns, S. J. & Lachniet, M. Multidecadal to multicentury scale collapses of Northern Hemisphere monsoons over the past millennium. *Proc. Natl Acad. Sci. USA* **110**, 9651–9656 (2013).
66. Myers, C. G. *et al.* Northeast Indian stalagmite records Pacific decadal climate change: implications for moisture transport and drought in India. *Geophys. Res. Lett.* **42**, 4124–4132 (2015).
67. Cook, E. R. *et al.* Asian monsoon failure and megadrought during the last millennium. *Science* **328**, 486–489 (2010).
68. Shen, M. *et al.* Evaporative cooling over the Tibetan Plateau induced by vegetation growth. *Proc. Natl Acad. Sci. USA* **112**, 9299–9304 (2015).
69. Cook, B. I. *et al.* Pre-Columbian deforestation as an amplifier of drought in Mesoamerica. *Geophys. Res. Lett.* **39**, L16706 (2012).
70. Fu, C. Potential impacts of human-induced land cover change on East Asia monsoon. *Global Planet. Change* **37**, 219–229 (2003).
71. Niyogi, D., Kishitwal, C., Tripathi, S. & Govindaraju, R. S. Observational evidence that agricultural intensification and land use change may be reducing the Indian summer monsoon rainfall. *Water Resour. Res.* **46**, W03533 (2010).
72. Yamashita, R., Matsumoto, J., Takata, K. & Takahashi, H. G. Impact of historical land-use changes on the Indian summer monsoon onset. *Int. J. Climatol.* **35**, 2419–2430 (2015).
73. Devaraju, N., Bala, G. & Modak, A. Effects of large-scale deforestation on precipitation in the monsoon regions: remote versus local effects. *Proc. Natl Acad. Sci. USA* **112**, 3257–3262 (2015).
- This model study suggests that large-scale deforestation results in a southward shift of the ITCZ and a decrease in monsoon rainfall in the Northern Hemisphere.**
74. Charney, J. G. Dynamics of deserts and drought in the Sahel. *Q. J. R. Meteorol. Soc.* **101**, 193–202 (1975).
75. Kröpelin, S. *et al.* Climate-driven ecosystem succession in the Sahara: the past 6000 years. *Science* **320**, 765–768 (2008).
76. Shanahan, T. M. *et al.* The time-transgressive termination of the African Humid Period. *Nature Geosci.* **8**, 140–144 (2015).
77. Liu, Z. *et al.* Simulating the transient evolution and abrupt change of Northern Africa atmosphere–ocean–terrestrial ecosystem in the Holocene. *Quat. Sci. Rev.* **26**, 1818–1837 (2007).
78. Rachmayani, R., Prange, M. & Schulz, M. North African vegetation–precipitation feedback in early and mid-Holocene climate simulations with CCSM3-DGVM. *Clim. Past* **11**, 175–185 (2015).
79. Swann, A. L. S., Fung, I. Y., Liu, Y. & Chiang, J. C. H. Remote vegetation feedbacks and the mid-Holocene green Sahara. *J. Clim.* **27**, 4857–4870 (2014).
80. Gray, L. J. *et al.* Solar influences on climate. *Rev. Geophys.* **48**, RG4001 (2010).
- This review discusses the contribution of solar variation to monsoon and North Atlantic climate change on decadal-to-centennial timescales.**
81. Schmidt, G. A. *et al.* Using palaeo-climate comparisons to constrain future projections in CMIP5. *Clim. Past* **10**, 221–250 (2014).
82. van Loon, H. & Meehl, G. A. The Indian summer monsoon during peaks in the 11 year sunspot cycle. *Geophys. Res. Lett.* **39**, L13701 (2012).
83. Zhao, L. & Wang, J.-S. Robust response of the East Asian monsoon rainband to solar variability. *J. Clim.* **27**, 3043–3051 (2014).
84. Wang, Y. *et al.* The Holocene Asian monsoon: links to solar changes and North Atlantic climate. *Science* **308**, 854–857 (2005).
85. Fleitmann, D. *et al.* Holocene forcing of the Indian monsoon recorded in a stalagmite from southern Oman. *Science* **300**, 1737–1739 (2003).
86. Meehl, G. A., Washington, W. M., Wigley, T. M. L., Arblaster, J. M. & Dai, A. Solar and greenhouse gas forcing and climate response in the twentieth century. *J. Clim.* **16**, 426–444 (2003).
87. Steinke, S. *et al.* Mid- to late-Holocene Australian–Indonesian summer monsoon variability. *Quat. Sci. Rev.* **93**, 142–154 (2014).
88. Sigl, M. *et al.* Timing and climate forcing of volcanic eruptions for the past 2,500 years. *Nature* **523**, 543–549 (2015).
89. Winter, A. *et al.* Persistent drying in the tropics linked to natural forcing. *Nature Commun.* **6**, 7627 (2015).
90. Ridley, H. E. *et al.* Aerosol forcing of the position of the intertropical convergence zone since ad 1550. *Nature Geosci.* **8**, 195–200 (2015).
91. Man, W., Zhou, T. & Jungclaus, J. H. Effects of large volcanic eruptions on global summer climate and East Asian monsoon changes during the last millennium: analysis of MPI-ESM simulations. *J. Clim.* **27**, 7394–7409 (2014).
92. Anchukaitis, K. J. *et al.* Influence of volcanic eruptions on the climate of the Asian monsoon region. *Geophys. Res. Lett.* **37**, L22703 (2010).
93. Park, J.-Y., Bader, J. & Matei, D. Northern-hemispheric differential warming is the key to understanding the discrepancies in the projected Sahel rainfall. *Nature Commun.* **6**, 5985 (2015).
94. Mohtadi, M. *et al.* Glacial to Holocene swings of the Australian–Indonesian monsoon. *Nature Geosci.* **4**, 540–544 (2011).
95. Collins, J. A. *et al.* Estimating the hydrogen isotopic composition of past precipitation using leaf-waxes from western Africa. *Quat. Sci. Rev.* **65**, 88–101 (2013).
96. Bollasina, M. A., Ming, Y. & Ramaswamy, V. Anthropogenic aerosols and the weakening of the South Asian summer monsoon. *Science* **334**, 502–505 (2011).
97. Yuan, Y. & Yang, S. Impacts of different types of El Niño on the East Asian climate: focus on ENSO cycles. *J. Clim.* **25**, 7702–7722 (2012).
98. Bordoni, S. & Schneider, T. Monsoons as eddy-mediated regime transitions of the tropical overturning circulation. *Nature Geosci.* **1**, 515–519 (2008).
99. Gadgil, S. The Indian monsoon and its variability. *Annu. Rev. Earth Planet. Sci.* **31**, 429–467 (2003).
100. Zhang, L. & Wang, C. Multidecadal North Atlantic sea surface temperature and Atlantic meridional overturning circulation variability in CMIP5 historical simulations. *J. Geophys. Res. Oceans* **118**, 5772–5791 (2013).

**Acknowledgements** We are grateful to J. H. C. Bosmans for providing Fig. 3. We thank F. He, Z. Liu and B. Otto-Bliesner for making the TraCE-21k model output available via the Earth System Grid (National Center for Atmospheric Research). This study is supported by the DFG Research Centre/Cluster of Excellence ‘The Ocean in the Earth System’ and the German Ministry of Education and Research (BMBF) grants 03G0228A (EISPAC), 03G0828A (TransGeoBioC) and 03G0484A (INVERS).

**Author Contributions** All authors determined the scope and wrote the paper, and contributed to interpretation and discussion of the results.

**Author Information** Reprints and permissions information is available at [www.nature.com/reprints](http://www.nature.com/reprints). The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to M.M. (mmohtadi@marum.de).



# The Atlantic salmon genome provides insights into rediploidization

Sigbjørn Lien<sup>1</sup>, Ben F. Koop<sup>2</sup>, Simen R. Sandve<sup>1</sup>, Jason R. Miller<sup>3</sup>, Matthew P. Kent<sup>1</sup>, Torfinn Nome<sup>1</sup>, Torgeir R. Hvidsten<sup>4,5</sup>, Jong S. Leong<sup>2</sup>, David R. Minkley<sup>2</sup>, Aleksey Zimin<sup>6</sup>, Fabian Grammes<sup>1</sup>, Harald Grove<sup>1</sup>, Arne Gjuvsland<sup>1</sup>, Brian Walenz<sup>3</sup>, Russell A. Hermansen<sup>7,8,9</sup>, Kris von Schalburg<sup>2</sup>, Eric B. Rondeau<sup>3</sup>, Alex Di Genova<sup>10,11</sup>, Jeevan K. A. Samy<sup>1</sup>, Jon Olav Vik<sup>1</sup>, Magnus D. Vigeland<sup>12</sup>, Lis Caler<sup>3</sup>, Unni Grimholt<sup>13</sup>, Sissel Jentoft<sup>14</sup>, Dag Inge Våge<sup>1</sup>, Pieter de Jong<sup>15</sup>, Thomas Moen<sup>16</sup>, Matthew Baranski<sup>17</sup>, Yniv Palti<sup>18</sup>, Douglas R. Smith<sup>19,20</sup>, James A. Yorke<sup>6</sup>, Alexander J. Nederbragt<sup>14</sup>, Ave Tooming-Klunderud<sup>14</sup>, Kjetill S. Jakobsen<sup>14</sup>, Xuanting Jiang<sup>21</sup>, Dingding Fan<sup>21</sup>, Yan Hu<sup>21</sup>, David A. Liberles<sup>8,9</sup>, Rodrigo Vidal<sup>22</sup>, Patricia Iturra<sup>23</sup>, Steven J. M. Jones<sup>24,25</sup>, Inge Jonassen<sup>26</sup>, Alejandro Maass<sup>10,11</sup>, Stig W. Omholt<sup>27</sup> & William S. Davidson<sup>25</sup>

**The whole-genome duplication 80 million years ago of the common ancestor of salmonids (salmonid-specific fourth vertebrate whole-genome duplication, Ss4R) provides unique opportunities to learn about the evolutionary fate of a duplicated vertebrate genome in 70 extant lineages. Here we present a high-quality genome assembly for Atlantic salmon (*Salmo salar*), and show that large genomic reorganizations, coinciding with bursts of transposon-mediated repeat expansions, were crucial for the post-Ss4R rediploidization process. Comparisons of duplicate gene expression patterns across a wide range of tissues with orthologous genes from a pre-Ss4R outgroup unexpectedly demonstrate far more instances of neofunctionalization than subfunctionalization. Surprisingly, we find that genes that were retained as duplicates after the teleost-specific whole-genome duplication 320 million years ago were not more likely to be retained after the Ss4R, and that the duplicate retention was not influenced to a great extent by the nature of the predicted protein interactions of the gene products. Finally, we demonstrate that the Atlantic salmon assembly can serve as a reference sequence for the study of other salmonids for a range of purposes.**

The 22,000-year-old cave painting of an Atlantic salmon (*Salmo salar*) near the Vézère River in France is a reminder of our fascination with, and dependence on, Atlantic salmon throughout human history. Atlantic salmon belongs to the salmonid lineage which comprises 11 genera, with at least 70 species that exhibit a wide range of ecological adaptations and use a variety of marine and freshwater life history strategies<sup>1</sup>. Salmonids hold important positions as socially iconic species and economic resources within aquaculture, wild fisheries and recreational sport fisheries. Moreover, they serve as key indicator species of the health of North Atlantic and Pacific coastal and river ecosystems.

All teleosts share at least three rounds of whole-genome duplication (WGD), 1R and 2R before the divergence of lamprey from the jawed vertebrates<sup>2</sup>, and a third teleost-specific WGD (Ts3R) at the base of the teleosts ~320 million years ago (Mya)<sup>3–5</sup>. Very little is known about the mechanisms of genomic and chromosomal reorganization after WGD in vertebrates because the 1R, 2R and Ts3R occurred so long ago that few clear signatures of post-WGD reorganization events remain. In contrast, a fourth WGD (the Ss4R salmonid-specific autotetraploidization event) occurred in the common ancestor of salmonids ~80 Mya after their divergence from Esociformes ~125 Mya<sup>6–8</sup> (Fig. 1), and the continued presence of multivalent pairing at meiosis and evidence of tetrasomic

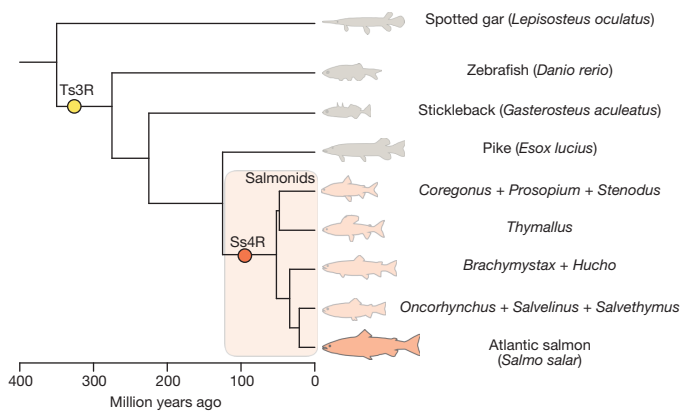
inheritance in salmonid species suggests that diploidy is not yet fully re-established<sup>6,9,10</sup>. Salmonids thus appear to provide an unprecedented opportunity for studying vertebrate genome evolution after an autotetraploid WGD<sup>11,12</sup> over a time period that is long enough to reveal long-term evolutionary patterns, but short enough to give a high-resolution picture of the process. In addition, they provide an excellent setting for contextualizing genome evolution with a dramatic post-WGD species radiation and intricate adaptations to a whole range of life history regimes.

Here we present a high-quality reference genome assembly of the Atlantic salmon, and use it to describe major patterns characterizing the post-Ss4R salmonid genome evolution over the past 80 million years (Myr). Our results challenge the recent claim that rediploidization in salmonids has been a gradual process unlinked to significant genome rearrangements<sup>13</sup>. They also challenge current views about the relative importance of sub- and neofunctionalization in vertebrate genomes (reviewed in ref. 14), and the importance of dosage balance as a gene duplicate retention mechanism<sup>15</sup>.

## Genome characterization

The Atlantic salmon reference genome assembly (GenBank: GCA\_000233375.4) adds up to 2.97 gigabases (Gb) with a

<sup>1</sup>Centre for Integrative Genetics (CIGENE), Department of Animal and Aquacultural Sciences, Norwegian University of Life Sciences, Ås NO-1432, Norway. <sup>2</sup>Department of Biology, University of Victoria, Victoria, British Columbia V8W 3N5, Canada. <sup>3</sup>J. Craig Venter Institute, 9704 Medical Center Drive, Rockville, Maryland 20850, USA. <sup>4</sup>Department of Chemistry, Biotechnology and Food Science, Norwegian University of Life Sciences, Ås NO-1432 Norway. <sup>5</sup>Department of Plant Physiology, Umeå Plant Science Centre, Umeå University, Umeå 90187, Sweden. <sup>6</sup>Institute for Physical Sciences and Technology, University of Maryland, College Park, Maryland 20742-2431, USA. <sup>7</sup>Department of Molecular Biology, University of Wyoming, Laramie, Wyoming 82071, USA. <sup>8</sup>Center for Computational Genetics and Genomics, Temple University, Philadelphia, Pennsylvania 19122-6078, USA. <sup>9</sup>Department of Biology, Temple University, Philadelphia, Pennsylvania 19122-6078, USA. <sup>10</sup>Center for Mathematical Modeling, University of Chile, Santiago 8370456, Chile. <sup>11</sup>Center for Genome Regulation, University of Chile, Santiago 8370415, Chile. <sup>12</sup>Medical Genetics, Oslo University Hospital and University of Oslo, Oslo NO-0424, Norway. <sup>13</sup>Department of Virology, Norwegian Veterinary Institute, Oslo NO-0454, Norway. <sup>14</sup>Centre for Ecological and Evolutionary Synthesis (CEES), Department of Biosciences, University of Oslo, Oslo NO-0316, Norway. <sup>15</sup>CHORI, Oakland, California 94609, USA. <sup>16</sup>AquaGen, Trondheim NO-7462, Norway. <sup>17</sup>Nofima, Tromsø NO-9291, Norway. <sup>18</sup>National Center for Cool and Cold Water Aquaculture, ARS-USDA, Kearneysville, West Virginia 25430, USA. <sup>19</sup>Beckman Genomics, Danvers, Massachusetts 01923, USA. <sup>20</sup>Courtagen Life Sciences, Woburn, Massachusetts 01801, USA. <sup>21</sup>BGI-Shenzhen, Shenzhen 518083, China. <sup>22</sup>Laboratory of Molecular Ecology, Genomics, and Evolutionary Studies, Department of Biology, University of Santiago, Santiago 9170022, Chile. <sup>23</sup>Faculty of Medicine, University of Chile, Santiago 8380453, Chile. <sup>24</sup>Genome Sciences Centre, BC Cancer Agency, Vancouver, British Columbia V5Z 4S6, Canada. <sup>25</sup>Department of Molecular Biology and Biochemistry, Simon Fraser University, Burnaby, British Columbia V5A 1S6, Canada. <sup>26</sup>Department of Informatics, University of Bergen, Bergen NO-6020, Norway. <sup>27</sup>Centre for Biodiversity Dynamics, Department of Biology, NTNU - Norwegian University of Science and Technology, Trondheim NO-7491, Norway.



**Figure 1 | Phylogenetic relationship of salmonids and relevant teleost lineages.** Divergence ages for salmonids are taken from ref. 8 and older divergences from ref. 7. *Parahucho* is not included in the figure due to uncertainty of its phylogenetic position. Ages do not represent the exact point estimates from the respective studies. Yellow and red circles represent the teleost specific whole genome duplication (Ts3R) and salmonid-specific whole genome duplication (Ss4R), respectively.

ctgN50 = 57.6 kb, which is consistent with genome size estimates<sup>16</sup>. Linkage mapping was used to position and orient 9,447 scaffolds (scfN50 = 2.97 megabases (Mb)), representing 2.24 Gb, into 29 single chromosome sequences (Supplementary Table 4). Most scaffolds not anchored to chromosomes consist of repetitive sequences. The 58–60% repeat content of Atlantic salmon is among the highest found in any vertebrate<sup>17</sup>. The single largest class of transposable elements is the Tc1-*mariner* family, representing 12.89% of the genome (Supplementary Information section 3). Tc1-*mariner* transposons tend to occur in centromeric regions (Fig. 2, track c), as reported in other species<sup>18</sup>.

Annotation of gene structures using RNA sequencing (RNA-seq) and expressed sequence tags (ESTs) identified 46,598 genes classified as non-repeat associated loci with sequence similarity support from the PFAM database, and/or zebrafish and stickleback annotations (Supplementary Table 11). Functional annotation identified a final set of 37,206 high-confidence protein-coding gene loci that have been assigned a putative functional annotation based on homology within the SwissProt database. Ninety-five per cent of the 498,245 public ESTs, and 98.3% of the identified loci were mapped to the 29 chromosome sequences, indicating a nearly complete representation of the protein-coding genome (Supplementary Information section 1.5).

### Post-Ss4R rediploidization characteristics

The return of a duplicated genome from tetrasomic to disomic inheritance relies on the obstruction of quadrivalent pairing during meiotic cell division. Large chromosome rearrangements through chromosome fusions, fissions, deletions or inversions strongly disrupt the possibility for homeologous pairing (the pairing of homeologue duplicates arising from a WGD)<sup>19,20</sup>. As extensive collinear blocks that include the telomere for at least one of the chromosome pairs is a diagnostic for current or recent multivalent pairing due to sequence homogenization (reviewed in ref. 21), we predicted that there would be an inverse relationship between homeologous sequence similarity and chromosome rearrangements in the duplicated blocks.

To test this prediction, we identified and analysed 98 homeologous (duplicated) blocks with high collinearity by aligning Atlantic salmon chromosome sequences against each other (Supplementary Information section 2). The 98 blocks (196 regions) account for 2.11 Gb (94.4%) of chromosome-anchored sequence (Fig. 2, Supplementary Table 6). A large proportion of homeologous blocks, representing roughly 573 Mb (25.6% of the chromosome-positioned

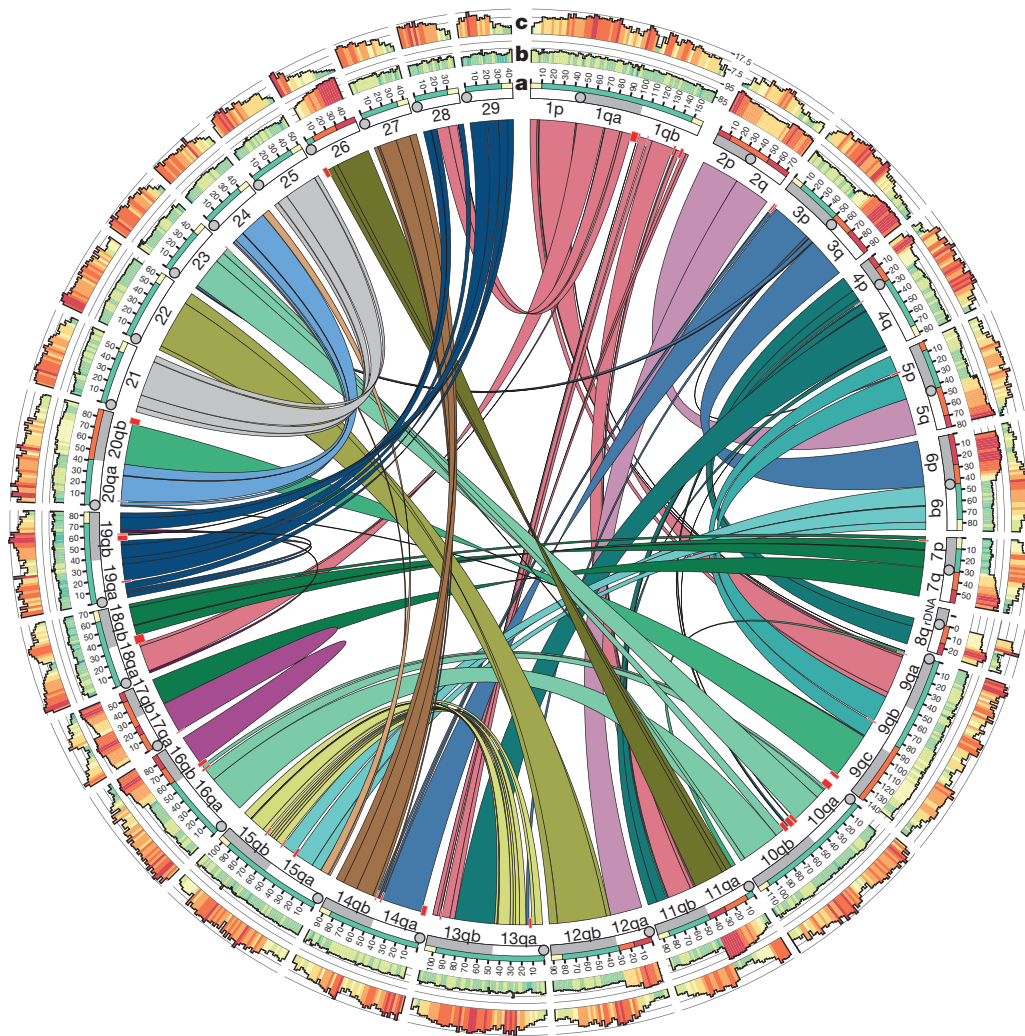
sequence), had a sequence similarity >90%. These regions were clustered within seven pairs of chromosome arms (2p–5q, 2q–12qa, 3q–6p, 4p–8q, 7q–17qb, 11qa–26, 16qb–17qa, and to some extent 9qc–20qb and 5p–9qb (Fig. 2)), and are all characterized by large collinear blocks including the telomere within at least one of the chromosome pairs. Previous studies in salmonids have claimed that at least one metacentric chromosome must be involved to provide the stability required for the formation of multivalents and homeologous pairing<sup>22</sup>. Our findings for regions 11qa–26 and 16qb–17qa indicate that this is not a strict necessity. Notably, increased read alignment depth and shorter scaffolds were characteristic of regions exceeding 95% similarity, representing 210 Mb (9.4% of the chromosome-positioned sequence), suggesting assembly collapse (Fig. 2, Supplementary Information section 1.5).

Without exception, duplicated regions exhibiting rearrangements at telomeres in the form of inversions, translocations or larger deletions all displayed a sequence similarity of ~87%. This clear correspondence between the degree of intra-block sequence similarity and blocks predicted to still participate in tetrasomic inheritance (or recently have done so) suggests that up to 25% of the salmon genome experienced delayed rediploidization after the initial large chromosome rearrangements, and that as much as 10% of the genome may still retain residual tetrasomy (Supplementary Table 7). The large and highly collinear blocks of shared synteny between Atlantic salmon and rainbow trout (Extended Data Fig. 1) imply that these rearrangements must have taken place before the split of the two lineages. This is also supported by combined genome mapping and karyotyping studies in other members of the Salmoninae subfamily, documenting conservation of large blocks embracing whole chromosome arms<sup>22</sup>.

To scrutinize this further, we analysed a set of 2,487 gene trees from orthologous gene sets containing putative homeologous pairs for both Atlantic salmon and rainbow trout (*Oncorhynchus mykiss*) (Supplementary Information section 5). As this analysis required calibration against an outgroup, we included only homeologous pairs having an orthologue in the Northern pike (*Esox lucius*), a member of the closest related diploid sister-group to salmonids<sup>23</sup>. Our results suggest ~100–80 Mya as a lower boundary for the Ss4R and that the *Salmo*–*Oncorhynchus* divergence occurred ~21 Mya (Fig. 3b; Extended Data Fig. 2c and Supplementary Information section 6), in agreement with recent age estimates<sup>8,13</sup>. Interestingly, analysis of asymmetry in coding sequence evolution between homeologues showed that a major part of the sequence divergence happened since the *Salmo*–*Oncorhynchus* split, suggesting a considerable temporal decoupling between the Ss4R event and sequence divergence of the Ss4R duplicates (Supplementary Information section 6). Moreover, our molecular dating results suggest that the majority of the Ss4R duplicates returned to disomic inheritance in a common ancestor of all salmonids before ~60 Mya (Fig. 3c). The results from the gene tree analysis are thus consistent with the data on homeologous sequence similarity (Extended Data Fig. 2b), strongly suggesting that large genomic reorganizations have been instrumental for the rediploidization process following the Ss4R. Our findings thus challenge one of the main conclusions from the recent sequencing of the rainbow trout genome, which suggested that rediploidization in salmonids has been a gradual process unlinked to significant genome rearrangements<sup>13</sup>.

Considering possible mechanisms underlying these large genomic reorganizations, the distribution of major transposon families in the Atlantic salmon genome suggests transposable element expansion in an ancestral salmonid before the chromosome fusions occurring in the Atlantic salmon lineage (Fig. 2, track c). The 85% sequence divergence among a large number of transposon family members is comparable to the lower boundary of homeologue block similarity (~87%). Assuming comparable neutral clock-like sequence divergence, this correspondence is consistent with a burst of repeat expansions coinciding with the initiation of rediploidization post-Ss4R (Fig. 3a and





**Figure 2 | The duplicated Atlantic salmon genome.** Homeologous regions in the Atlantic salmon genome subdivided into 98 collinear blocks along the 29 European Atlantic salmon chromosomes. Red rectangles represent blocks of sequence without identifiable duplicated regions elsewhere in the genome. **a**, This track shows grouping of salmon sequence into regions; red = high (>95% sequence similarity), orange = elevated

(90–95% sequence similarity), green = low (~87% sequence similarity), yellow = telomeric regions (10 Mb) characterized by highly elevated male recombination (see ref. 10). **b**, This track shows genomic similarity (in 1 Mb intervals) between duplicated regions (red = high, yellow = medium, green = low sequence similarity). **c**, This track shows frequency of Tc1-mariner transposon elements in the Atlantic salmon genome.

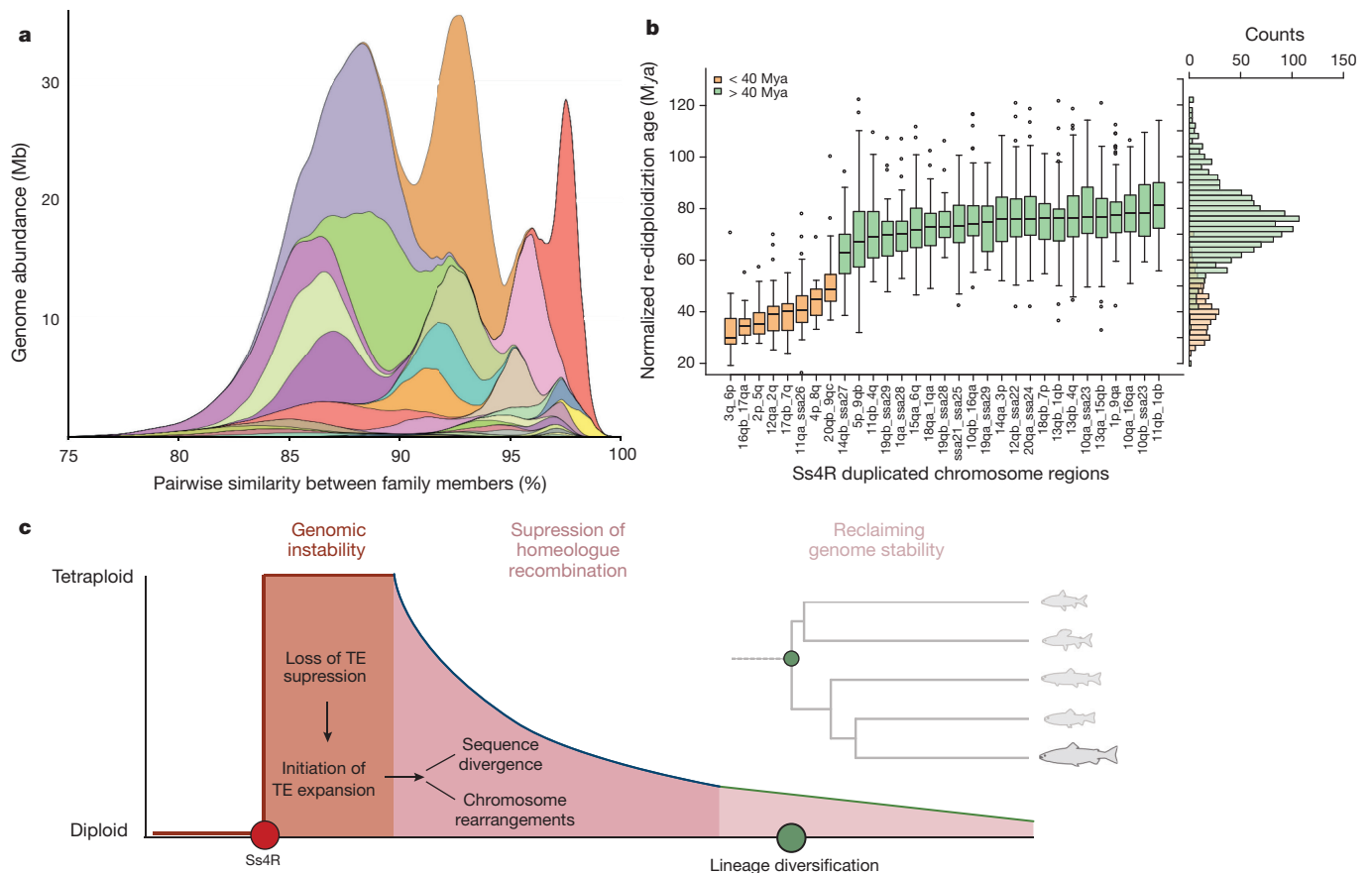
Extended Data Fig. 2b and Supplementary Information section 6.2). As large-scale expansion and movement of transposable elements are known to increase under genomic stress<sup>24</sup>, this may suggest that Ss4R caused transposable element expansion by compromising regulatory processes responsible for transposon policing. This expansion might in turn have been a major determinant for driving the genome towards a diploid state through enhanced homeologue sequence divergence and large chromosome rearrangements due to ectopic transposable element recombination and chromosomal breakage causing non-homologous end-joining<sup>25</sup> (Fig. 3c).

### Duplicate retention—patterns and mechanisms

To assess the evolutionary fates of duplicated genes in the salmon genome, we analysed patterns of Ss4R duplicate retention and functional divergence of protein-coding genes within the 98 homeologous blocks. Considering that we find very little evidence for gene loss through fractionation<sup>26</sup>, and that in 56% of the 9,162 singletons we were able to identify a pseudogenized homeologue gene fragment in an expected position (Supplementary Information section 4 and Supplementary Table 11), pseudogenization appears to be the predominant mechanism underlying Ss4R duplicate loss.

To contrast the Ss4R with the 240 Myr older Ts3R duplicate retention patterns, we analysed duplicate retention patterns in teleost gene family trees (ref. 27; Supplementary Information section 8). This revealed that 20% of the Ts3R and 55% of the Ss4R duplicates are retained as two functional copies in Atlantic salmon. In comparison, 12–24% of duplicated genes derived from the Ts3R event have been retained in other extant teleost fish lineages (reviewed in ref. 28), and the retention 75 Myr post-Ts3R has been estimated to have been about 40%<sup>3,29</sup>. Considering the uncertainty attached to such estimates, the post-Ss4R temporal retention profile of Atlantic salmon is arguably quite similar to that of other teleosts post-Ts3R, indicating that mechanisms responsible for duplicate retention in Atlantic salmon may be generic.

Surprisingly, Atlantic salmon genes that were retained as duplicates after the Ts3R event were not more likely to be retained after the Ss4R (Extended Data Fig. 3; Supplementary Information section 8). The predominantly independent probabilities of retention suggest a complex interplay of processes, different evolutionary drivers of duplicate retention, or a largely neutral and stochastic nonfunctionalization process following the Ts3R and Ss4R events. Interestingly, we observed enhanced retention of non-WGD gene duplicates (older or younger than the Ss4R



**Figure 3 | Post-Ss4R rediploidization.** **a**, Fig. 3a shows a significant and ongoing expansion of transposable elements from the Tc1-mariner superfamily with major peaks at an average of 87%, 93% and 98% similarity between family members. The colours correspond to the same colours as in the box plot in Extended Data Fig. 5. **b**, Age estimates of the

time from homeologue divergence to *Salmo-Oncorhynchus* divergence for each individual homeologous region. Only chromosome regions with >10 gene trees were included. **c**, A three-step hypothetical model of post-Ss4R rediploidization (widths of model compartments do not reflect actual time scales). The green circle indicates the beginning of the salmonid radiation.

event) when the WGD (both Ts3R and Ss4R) duplicates also had been retained ( $P < 0.001$ ; Supplementary Information section 8).

Two major mechanisms by which a pair of duplicates can escape the fate of nonfunctionalization are subfunctionalization (partitioning of ancestral gene functions)<sup>30</sup> and neofunctionalization (assigning a novel function to one of the duplicates)<sup>31</sup>. To assess the relative importance of these two mechanisms we analysed gene expression divergence of Ss4R duplicates across 15 tissues (Extended Data Fig. 4a, b; Supplementary Information section 7). Forty-five per cent (3,991/8,954) of well-defined expressed Ss4R pairs showed signs of diverged expression by being located in different co-expression clusters (Fig. 4a). Diverged homeologues tended to belong to closely related but still different co-expression clusters (Fig. 4a and Extended Data Fig. 4d).

Although these results suggest that functional divergence is common among Ss4R duplicates, information about ancestral state is critical for the classification of this divergence into sub- and neofunctionalization. We therefore used comparable expression data across 13 common tissues from diploid Northern pike<sup>23</sup> as a proxy for the ancestral state of Ss4R duplicates. We identified 8,102 orthologous gene triplets (that is, two Ss4R copies and their putative pike orthologue) and in 42% of the triplets both Ss4R duplicates showed conserved co-expression profile with the pike orthologue (Pearson correlation  $> 0.6$ ,  $P < 0.03$ ). This indicates strong purifying selection pressure on gene regulation across more than 100 Myr and adds credibility to the use of Northern pike for assessing ancestral gene regulation. In 28% of the triplets, one Ss4R duplicate had a conserved co-expression pattern

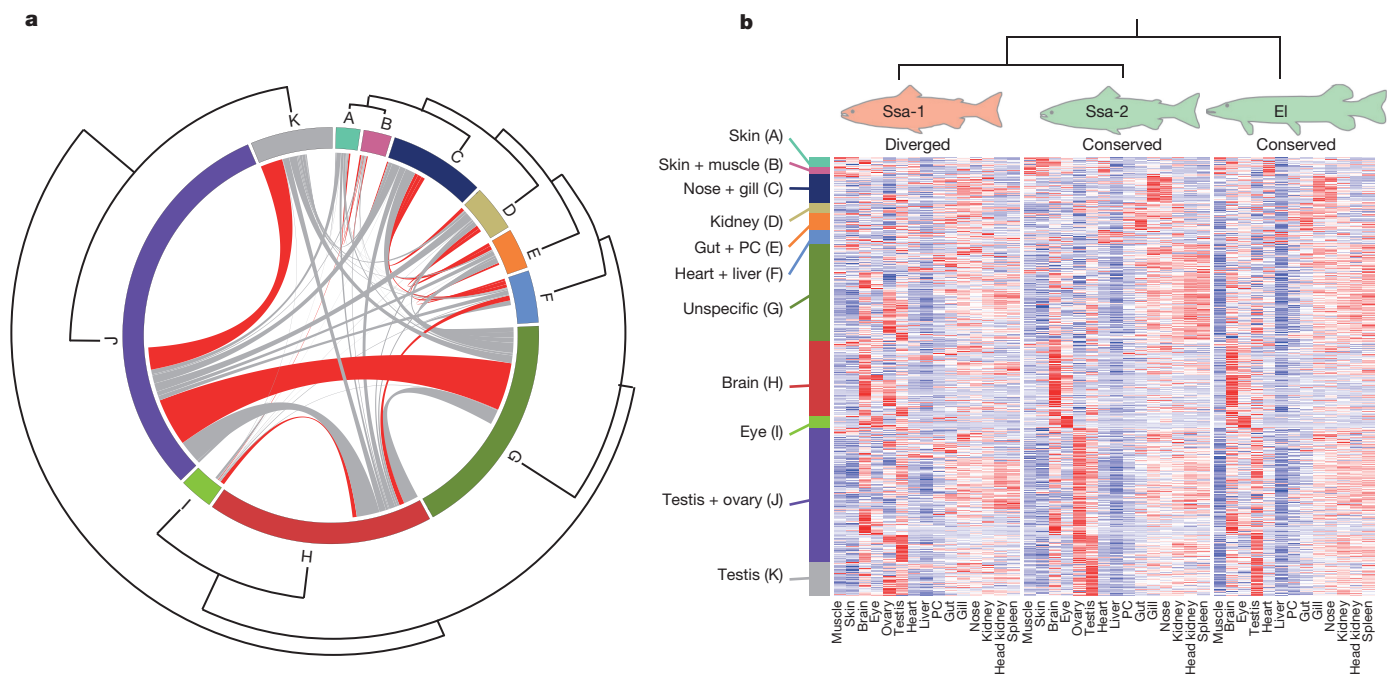
with pike and the other belonged to a different co-expression cluster (Fig. 4b), indicative of regulatory neofunctionalization.

Although we observed cases of putative pseudogenization in Ss4R duplicates displaying a low correlation in expression regulation in combination with large coding sequence length difference, most Ss4R duplicates had similar lengths regardless of their expression similarity (Extended Data Fig. 4e), suggesting that neutral evolution can only marginally explain this regulatory divergence.

We identified 1,084 triplets where the salmon duplicates belonged to different expression clusters and had expression profiles significantly different from pike (Pearson correlation  $< 0.55$ ,  $P > 0.05$ ), pointing to possible subfunctionalization. In this group we found, somewhat surprisingly, only 23 clear examples of subfunctionalization where the sum of the expression patterns of salmon homeologues correlated significantly with assumed ancestral state. However, this cluster-based analysis neglects subtler within-cluster subfunctionalization cases, as well as those involving acquisition of novel functions after subfunctionalization. To account for this, we applied an 'on-off' classification method (Extended Data Fig. 4f and Supplementary Information section 7.2) that increased the estimate to 167 cases; a figure that is still dwarfed by the estimated number of neofunctionalization cases (3,028) (Supplementary Information section 7.2).

Purifying selection on dosage sensitive interactions with other duplicated genes is thought to be an important mechanism for intermediate duplicate retention after WGDs<sup>15</sup>, before neo-, sub- and nonfunctionalization determine the ultimate fate of the duplicates<sup>32</sup>. In line with this, we observed an overrepresentation of GO terms





**Figure 4 | Homeologue divergence.** **a**, Circos plot distribution of homeologous gene pairs and their assignment to 11 co-expression clusters based on 15 different tissues. Lines connect Ss4R pairs that belong to different co-expression clusters. For visualization purposes, we sorted the Ss4R pairs according to type of co-expression divergence. Red lines signify

associated with signal transduction, protein complex formation and transcription among the duplicated genes with conserved regulation (Supplementary Information section 7.3 and Supplementary Table 16). However, as a diversity of GO terms not focal to the dosage balance hypothesis (Supplementary Table 16) are also overrepresented among Ss4R duplicates with conserved regulation, it is not justified to conclude that dosage balance is the sole intermediate retention mechanism. Furthermore, analyses of retention patterns after Ts3R and Ss4R suggest independent retention probability and a very weak effect of preferential co-retention of known protein interacting partners ( $P < 0.001$ ) for both the Ts3R and Ss4R duplication events (Extended Data Fig. 3 and Supplementary Information section 8).

Taken together, >60% of the homeologue pairs show signatures of tissue-dependent regulatory divergence at the whole gene or exon-level (Supplementary Information section 7.2). The predominance of cases where only one copy has changed its regulation compared to the assumed ancestral state indicates that regulatory subfunctionalization has not been a dominant duplicate retention mechanism post Ss4R, unless it was followed by subsequent neofunctionalization, which has been suggested as a common process<sup>33,34</sup>. However, our subfunctionalization estimates together with the high frequency of triplets where one salmon homeologue had a conserved co-expression pattern with pike while its duplicate did not (Fig. 4b), are not consistent with the generality of this latter scenario.

### A reference genome for salmonids

Conservation of synteny between salmonids<sup>22,35</sup> suggests that information from one high-quality salmonid genome can be used to improve genome sequence assemblies of other salmonids. To test the feasibility of such a comparative genomics approach, we used the Atlantic salmon assembly to construct chromosome sequences for the non-chromosome anchored rainbow trout genome sequence<sup>13</sup>. We were able to map 99.5% of rainbow trout scaffolds >100 kilobases (kb) (total 1.22 Gb) to the Atlantic salmon chromosome sequences (Supplementary Information section 1.5).

significant resampling tests ( $P < 0.05$ ) for enrichment of homeologue divergence between two specific co-expression clusters. **b**, Heatmap of 2,272 triplets (two salmon homeologues and a pike orthologue), in which one of the Atlantic salmon homeologues has diverged in gene expression

Using the Atlantic salmon chromosome sequences together with a dense linkage map for rainbow trout constructed from a 57K single nucleotide polymorphisms (SNP) array, we were able to anchor, orient and concatenate 11,335 rainbow trout scaffolds (scfN50 = 940 kb, from ref. 13) into 29 rainbow trout chromosome sequences (Supplementary Information section 9). This was done by first using the rainbow trout linkage map to determine the proximate order of 2,439 trout scaffolds containing SNPs, which we found to be sufficient for determining conserved blocks. Then we used comparative information from Atlantic salmon to incorporate scaffolds without SNP information, and fine-tune the order and orientation of all 11,335 trout scaffolds into chromosome sequences. Even though the rainbow trout linkage map contains more markers than most other salmonids (for example, ref. 22), this high number of properly placed scaffolds would not be achievable without the Atlantic salmon information.

Alignment of these rainbow trout chromosomes (representing 1.37 Gb of sequence) with the Atlantic salmon genome revealed conservation of very large syntenic blocks, in many cases corresponding to whole chromosome arms in rainbow trout (Extended Data Fig. 1). This analysis supports previous results<sup>35</sup> suggesting conservation of 50 syntenic regions representing the karyotype of 50 acrocentric chromosomes in the common ancestor of salmonids<sup>36</sup>. Our analysis documents that these syntenic regions typically represent blocks with no rearrangements for 38 regions and with only one or two inversions or translocations among the remaining parts.

### Implications

The conservation of large collinear blocks between *Salmo* and *Oncorhynchus* strongly suggests that the Atlantic salmon genome information will facilitate exploitation of genomic information in a wide range of ecological, evolutionary, conservation and production biology settings within salmonids. Moreover, the availability of a high-quality assembly and annotation of the Atlantic salmon genome provides novel insights into vertebrate post-WGD evolution that may contribute to a more thorough understanding of the underlying mechanisms as well as the long-term importance of WGD for adaptation.

**Online Content** Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

**Received 10 September 2015; accepted 26 January 2016.**

**Published online 18 April 2016.**


- Nelson, J. S. *Fishes of the World* (John Wiley & Sons, 2006).
- Smith, J. J. *et al.* Sequencing of the sea lamprey (*Petromyzon marinus*) genome provides insights into vertebrate evolution. *Nature Genet.* **45**, 415–421 (2013).
- Jaillon, O. *et al.* Genome duplication in the teleost fish *Tetraodon nigroviridis* reveals the early vertebrate proto-karyotype. *Nature* **431**, 946–957 (2004).
- Kasahara, M. *et al.* The medaka draft genome and insights into vertebrate genome evolution. *Nature* **447**, 714–719 (2007).
- Nakatani, Y., Takeda, H., Kohara, Y. & Morishita, S. Reconstruction of the vertebrate ancestral genome reveals dynamic genome reorganization in early vertebrates. *Genome Res.* **17**, 1254–1265 (2007).
- Allendorf, F. W. & Thorgaard, G. H. in *Evolutionary Genetics of Fishes* (ed. Turner, B. J.) 1–53 (Plenum Press, 1984).
- Near, T. J. *et al.* Resolution of ray-finned fish phylogeny and timing of diversification. *Proc. Natl Acad. Sci. USA* **109**, 13698–13703 (2012).
- Macqueen, D. J. & Johnston, I. A. A well-constrained estimate for the timing of the salmonid whole genome duplication reveals major decoupling from species diversification. *Proc. R. Soc. B* **281**, 20132881 (2014).
- Wright, J. E., Johnson, K., Hollister, A. & May, B. Meiotic models to explain classical linkage, pseudolinkage, and chromosome pairing in tetraploid derivative salmonid genomes. *Iszymes Curr. Top. Biol. Med. Res.* **10**, 239–260 (1983).
- Lien, S. *et al.* A dense SNP-based linkage map for Atlantic salmon (*Salmo salar*) reveals extended chromosome homeologies and striking differences in sex-specific recombination patterns. *BMC Genomics* **12**, 615 (2011).
- Davidson, W. S. *et al.* Sequencing the genome of the Atlantic salmon (*Salmo salar*). *Genome Biol.* **11**, 403 (2010).
- Mayfield-Jones, D. *et al.* Watching the grin fade: tracing the effects of polyploidy on different evolutionary time scales. *Semin. Cell Dev. Biol.* **24**, 320–331 (2013).
- Berthelot, C. *et al.* The rainbow trout genome provides novel insights into evolution after whole-genome duplication in vertebrates. *Nat. Commun.* **5**, 3657 (2014).
- Glasauer, S. M. K. & Neuhauss, S. C. F. Whole-genome duplication in teleost fishes and its evolutionary consequences. *Mol. Genet. Genomics* **289**, 1045–1060 (2014).
- Schnable, J. C., Pedersen, B. S., Subramaniam, S. & Freeling, M. Dose-sensitivity, conserved non-coding sequences, and duplicate gene retention through multiple tetraploidies in the grasses. *Front. Plant Sci.* **2**, 2 (2011).
- Hardie, D. C. & Hebert, P. D. N. The nucleotypic effects of cellular DNA content in cartilaginous and ray-finned fishes. *Genome* **46**, 683–706 (2003).
- McCluskey, B. M. & Postlethwait, J. H. Phylogeny of zebrafish, a “model species,” within *Danio*, a “model genus”. *Mol. Biol. Evol.* **32**, 635–652 (2015).
- Daron, J. *et al.* Organization and evolution of transposable elements along the bread wheat chromosome 3B. *Genome Biol.* **15**, 546 (2014).
- Wendel, J. F. Genome evolution in polyploids. *Plant Mol. Biol.* **42**, 225–249 (2000).
- Gerstein, A. C., Chun, H.-J. E., Grant, A. & Otto, S. P. Genomic convergence toward diploidy in *Saccharomyces cerevisiae*. *PLoS Genet.* **2**, e145 (2006).
- Allendorf, F. W. *et al.* Effects of crossovers between homeologs on inheritance and population genomics in polyploid-derived salmonid fishes. *J. Hered.* **106**, 217–227 (2015).
- Kodama, M., Briec, M. S. O., Devlin, R. H., Hard, J. J. & Naish, K. A. Comparative mapping between Coho salmon (*Oncorhynchus kisutch*) and three other salmonids suggests a role for chromosomal rearrangements in the retention of duplicated regions following a whole genome duplication event. *G3 Genes Genomes Genetics* **4**, 1717–1730 (2014).
- Rondeau, E. B. *et al.* The genome and linkage map of the Northern pike (*Esox lucius*): conserved synteny revealed between the salmonid sister group and the neoteleostei. *PLoS ONE* **9**, e102089 (2014).
- Slotkin, R. K. & Martienssen, R. Transposable elements and the epigenetic regulation of the genome. *Nature Rev. Genet.* **8**, 272–285 (2007).
- Guillén, Y. & Ruiz, A. Gene alterations at *Drosophila* inversion breakpoints provide prima facie evidence for natural selection as an explanation for rapid chromosomal evolution. *BMC Genomics* **13**, 53 (2012).
- Langham, R. J. *et al.* Genomic duplication, fractionation and the origin of regulatory novelty. *Genetics* **166**, 935–945 (2004).
- Cunningham, F. *et al.* Ensembl 2015. *Nucleic Acids Res.* **43**, D662–D669 (2015).
- Braasch, I. & Postlethwait, J. H. in *Polyploidy and Genome Evolution* (eds Soltis, P. S. & Soltis, D. E.) Polyploidy in fish and the teleost genome duplication (Springer, 2012).
- Sato, Y., Hashiguchi, Y. & Nishida, M. Temporal pattern of loss/persistence of duplicate genes involved in signal transduction and metabolic pathways after teleost-specific genome duplication. *BMC Evol. Biol.* **9**, 127 (2009).
- Force, A. *et al.* Preservation of duplicate genes by complementary, degenerative mutations. *Genetics* **151**, 1531–1545 (1999).
- Conant, G. C. & Wolfe, K. H. Turning a hobby into a job: how duplicated genes find new functions. *Nature Rev. Genet.* **9**, 938–950 (2008).
- Hughes, T., Ekman, D., Ardawatia, H., Elofsson, A. & Liberles, D. A. Evaluating dosage compensation as a cause of duplicate gene retention in *Paramecium tetraurelia*. *Genome Biol.* **8**, 213 (2007).
- He, X. & Zhang, J. Rapid subfunctionalization accompanied by prolonged and substantial neofunctionalization in duplicate gene evolution. *Genetics* **169**, 1157–1164 (2005).
- Rastogi, S. & Liberles, D. A. Subfunctionalization of duplicated genes as a transition state to neofunctionalization. *BMC Evol. Biol.* **5**, 28 (2005).
- Phillips, R. B. *et al.* Assignment of Atlantic salmon (*Salmo salar*) linkage groups to specific chromosomes: conservation of large syntenic blocks corresponding to whole chromosome arms in rainbow trout (*Oncorhynchus mykiss*). *BMC Genet.* **10**, 46 (2009).
- Mank, J. E. & Avise, J. C. Phylogenetic conservation of chromosome numbers in Actinopterygian fishes. *Genetica* **127**, 321–327 (2006).

**Supplementary Information** is available in the online version of the paper.

**Acknowledgements** The International Cooperation to Sequence the Atlantic Salmon Genome (ICSASG) was funded by the following organizations: Research Council of Norway (NFR; <http://www.rcn.no>), Norwegian Seafood Research Fund (<http://www.hfh.no/hot-topics/about-hfh>), Genome BC (<http://www.genomebc.ca>), The Chilean Economic Development Agency – CORFO and InnovaChile Committee (<http://www.english.corfo.cl>), Marine Harvest (<http://www.marineharvest.no>), AquaGen (<http://www.aquagen.no>), Cermaq (<http://www.cermaq.com>) and Salmobreed (<http://www.salmobreed.no>). Additional funding for bioinformatics and statistical support at CIGENE-NMBU was provided by NFR grants 208481/F50, 226266, 225181 and 221734/O30. Funding for RNA-seq and most of the repeat analysis was provided by Natural Sciences and Engineering Research Council (NSERC), Canada. NSERC also provided funding for the analyses and assemblies generated at University of Victoria throughout the project. We acknowledge the help of S. Karoliussen, M. Arnyasi, R. Martinsen Ånestad and I. Johansson Schneider at CIGENE-NMBU for generating salmon and rainbow trout genotypes, and G. Gao, National Center for Cool and Cold Water Aquaculture, ARS-USDA, for generating rainbow trout genotypes. We also acknowledge K. Beeson and H. Baden-Tilson at the J. Craig Venter Institute for library construction and Illumina sequencing. Bioinformatic analyses were performed using resources at the Orion Computing Cluster at CIGENE-NMBU, the Norwegian metacenter for computational science (under project nn4653k), computing resources at University of Victoria partly provided by Compute Canada, and NLHPC-Chile.

**Author Contributions** S.L., B.F.K., S.W.O. and W.S.D. conceived the study. The project was led by the Executive Scientific Committee (ESC) of the International Collaboration to Sequence the Atlantic Salmon Genome (ICSASG) consisting of: S.L., B.F.K., A.M., R.V., P.I., S.J.M.J., I.J., S.W.O. and W.S.D. U.G., D.I.V., S.J. and S.W.O. produced and nurtured the double haploid salmon. P.d.J. made the BAC library. B.F.K., M.P.K., J.R.M., L.C., D.R.S., A.T.-K., A.J.N., K.S.J., X.J., D.F. and Y.H. provided sequence data. J.R.M., B.W., A.Z., B.F.K., J.S.L., J.A.Y., A.D.G., A.J.N., T.N., H.G. and S.L. produced and refined the assembly. T.N., H.G., and S.L. built chromosome sequences. S.L., M.P.K., T.N., H.G., A.G., T.M., M.B. and Y.P. produced and analysed SNP data. S.R.S., B.F.K., F.G., T.N., T.R.H., J.K.A.S., J.S.L., D.R.M., K.v.S. and E.B.R. generated RNA data and completed gene annotation. B.F.K. and D.R.M. performed repeat element analyses. S.R.S., T.N., A.G., D.A.L., R.A.H. and S.W.O. performed analyses on post-Ss4R rediploidization. S.R.S., T.R.H. and M.D.V. carried out evolutionary and comparative analyses. S.L., T.N., H.G., B.F.K., J.S.L., K.v.S., E.R. T.M., M.B. and Y.P. produced data and completed comparative genome analyses. T.R.H., A.G., J.O.V. and M.D.V. performed additional statistical analyses. S.L., B.F.K., S.R.S., M.P.K., T.N., T.R.H., S.W.O. and W.S.D. wrote the manuscript. All authors read and commented on the manuscript.

**Author Information** Sequence information was deposited at GenBank under accession code GCA\_000233375.4 and at the NCBI Sequence Read Archive (SRA): PRJNA72713 and PRJNA260929. Reprints and permissions information is available at [www.nature.com/reprints](http://www.nature.com/reprints). The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to W.S.D. ([wdavidso@sfu.ca](mailto:wdavidso@sfu.ca)) or S.W.O. ([stigmoholt@ntnu.no](mailto:stigmoholt@ntnu.no)).

 This work is licensed under a Creative Commons Attribution 4.0 International (CC BY 4.0) licence. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in the credit line; if the material is not included under the Creative Commons licence, users will need to obtain permission from the licence holder to reproduce the material. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.



## METHODS

**Data reporting.** No statistical methods were used to predetermine sample size. The investigators were not blinded to allocation during experiments and outcome assessment.

**Genome sequencing and assembly.** DNA from a single double-haploid female from the AquaGen strain, produced by mitotic androgenesis, served as the template for sequencing using Sanger and next generation sequencing technologies (Supplementary Table 1). Various assemblies were generated using different combinations of software and subsets of data (Supplementary Table 2). The foundation of the chosen assembly was generated from Sanger ( $\sim 4\times$ ) and Illumina ( $\sim 202\times$ ) data assembled using the MaSuRCA (v2.0.3) assembler<sup>37</sup>. The assembly was reconciled and gap-filled using information from preliminary assemblies (Supplementary Information section 1.3). Genetic linkage information describing 565,877 SNPs was used to both confirm and correct scaffolds and, when supported by information from other assemblies, was used to join scaffolds within linkage groups. Subsequently, linkage analysis using CRIMAP<sup>38</sup> and a subset of SNP sequence tags (27,221) were used to order, orient and concatenate scaffolds into 29 single-chromosome sequences. Nomenclature for Atlantic salmon chromosomes is based on ref. 35.

**Gene annotation.** Gene structures were determined by combining data from full-length cDNA sequences<sup>39</sup>, EST databases<sup>39–41</sup>, and RNA-seq data from 15 tissues (Supplementary Table 9). RNA-seq reads were trimmed using Trimmomatic (v0.32 (ref. 42)) and mapped to the reference genome sequence using STAR (v2.3.1z12 (ref. 43)), and all publicly available mRNAs and ESTs were mapped using GMAP<sup>44</sup>. Gene structures were predicted with CUFLINKS<sup>45</sup>. Open reading frame (ORF) predictions were carried out using TransDecoder<sup>46</sup>. Gene models without homology match to either PFAM, stickleback or zebrafish were discarded. Functional annotation was done with Blast2GO<sup>47</sup> against the SwissProt database. Transposable element related ORFs were identified with BLAST searches against the annotated transposable element sequences and queries in the functional annotation gene names for transposable element related terms (that is, retrotransposon, transposon, transposase, transposase, reverse transcriptase, gag, bpol). Putative expressed and silenced Ss4R homeologues were identified using a combination of homology searches with BLAST and GenomeThreader<sup>48</sup> targeting a priori defined conserved collinear duplicated regions ( $n=98$ ).

**Repeat library methods.** An Atlantic salmon repeat library of 2,005 elements was assembled from sequences previously reported in salmonids<sup>13,49,50</sup> and the output of the *de novo* repeat-finding programs LTRharvest<sup>51</sup>, RepeatModeller<sup>52</sup> and REPET<sup>53</sup>. With the exception of curated repeats previously reported by Matveev and Okada<sup>50</sup> and those found in the RepBase database<sup>49</sup>, all preliminary sequences were validated using BLASTn<sup>54</sup> to ensure that they were present at multiple locations in the genome. LTRharvest sequences were filtered based on the repeat library construction procedure outlined in the MAKER documentation<sup>55</sup>. Using BLASTn, sequences from other *de novo* sources and the rainbow trout repeat library were flagged as potentially chimaeric if they did not generate at least three high-scoring segment pairs (HSPs) covering at least 80% of their length in the Atlantic salmon genome. Any distinct highly repetitive region within such sequences was extracted and retained while other portions were discarded. All libraries were merged and redundant sequences were removed based on the guidelines presented by Wicker *et al.*<sup>56</sup> and the MAKER documentation. Sequences in the combined library were annotated, and non-transposable element host genes were removed based on their similarity to well-characterized sequences in annotation databases<sup>49,57</sup>, the presence of structural motifs and manual examination.

To estimate the historical activity of Tc1-*mariner* transposable elements, up to 100 randomly selected full-length genomic copies from each of 40 Tc1-*mariner* families were extracted and aligned using MUSCLE<sup>58</sup>. All families were confirmed to be phylogenetically distinct from each other and possessed a star-like neighbour-joining tree topology characteristic of Tc1-*mariner* activity<sup>59</sup>. The distribution of pairwise per cent similarity, a proxy for time, between members of a family was used to analyse the temporal dynamics of transposable element activity. **Identification of homeologous blocks within the salmon genome.** Repeat masked chromosome sequences for Atlantic salmon (see above) were aligned against each other using LASTZ<sup>60</sup> to identify 98 homologous blocks originating from the Ss4R (for details see Supplementary Information section 2). Sequence similarity between homeologous sequences were determined in 1 Mb intervals by averaging local percentage of nucleotide sequence identity using high-scoring segment pair (HSP) from LASTZ alignments<sup>60</sup> and presented as a Circos plot<sup>61</sup> in Figure 2.

**Sequence evolution analyses of salmon homeologues.** Putative orthologue sequence sets were collated with Best Reciprocal Blast (BRB) protein matches. For salmonid species the top-two BRB-hits were assigned to putative orthologue groups. Multiple codon sequence alignments were constructed using MAFFT<sup>62</sup>

and quality trimmed with Guidance in an iterative framework where sequences were re-aligned after identification of poorly aligned codons.

Maximum likelihood (ML) gene trees were calculated by the R-package Phangorn<sup>63</sup> using codon alignments, the GTR+G+I model, and 100 bootstrap replicates. Branch specific GTR+G+I substitution rates were estimated functions from the R-package ape<sup>64</sup>, while branch specific synonymous (dS) and non-synonymous (dN) substitution rates were estimated with non-negative least squares regression in the Phangorn R package<sup>63</sup> using pairwise dN and dS distance matrixes from codeml<sup>65</sup> and the ML gene tree topologies as input.

Branch-site specific test for positive selection was carried out by a likelihood-ratio test on the ML-likelihood estimates for sequence evolution under different models in codeml. The smallest likelihood estimate from four omega starting values (0.5, 1, 1.5, and 2) was used in the likelihood ratio test (LRT). False discovery rate adjustments of p-values were done with the p.adjust function in R.

**Gene tree dating.** BEAST<sup>66</sup> was used to calibrate gene trees using a HKY+G substitution model, uncorrelated lognormal clock, and yule tree prior. The BEAST analyses were exclusively based on codon alignments that produced a ML-gene tree topology containing two Ss4R homeologues in both *Salmo* and *Oncorhynchus*, and where rediploidization had occurred before the *Salmo*–*Oncorhynchus* divergence. No priors on tree topology were specified and a single secondary calibration of 127 Myr (confidence interval 12.5 Myr) on the most recent common ancestor of *Salmoniformes* + *Esociformes* was used<sup>7,8</sup>. All Markov chain Monte Carlo (MCMC) analyses were run for 10 million generations with sampling every, 1000 generations. Tracer v1.6 (available from <http://beast.bio.ed.ac.uk/Tracer>) was used to inspect effective sample sizes (ESS) of tree parameters. Fifty per cent consensus topologies were constructed based on 100 randomly sampled tree topologies from the last 1,000 MCMC-samples. Age of *Salmo*–*Oncorhynchus* divergence was estimated as the median of two nodes per tree.

**Transcriptome analysis.** A gene was classified as ‘expressed’ if the FPKM value of at least one tissue was above 1.0, and values were transformed to  $\log_2(\text{FPKM}+1)$  values for consecutive analysis. Samples and genes were clustered using Pearson correlation and Ward’s method in the R function hclust<sup>67</sup>, and visualized as heatmaps using the R function heatmap.2 (gplots library). Genes were scaled individually in the heatmaps.

Clusters with a significant number of shared homeologue-pairs were identified by simulation (10,000 randomizations). A salmon gene (or exon) was classified as conserved if the Pearson correlation to the pike orthologue was above 0.6 ( $P=0.03$ ) across the 13 common tissues, and diverged if the correlation was below 0.55 ( $P>0.05$ ). A salmon homeologue-pair was classified as neofunctionalized if at least one salmon gene was conserved and the two salmon genes were in different clusters, and as subfunctionalized if both salmon genes were diverged and in different clusters, but their summed expression was conserved.

Expression specificity was computed as one minus the sum, over all samples, of the gene’s expression in that sample divided by the maximum expression in any sample. Significant difference in specificity between clusters was computed using the Wilcoxon test.

**Duplicate retention.** Existing gene families for all teleost species were downloaded from Ensembl Compara 79 (ref. 27). Genomes for *Salmo salar*, *Esox lucius*, and *Oncorhynchus mykiss* were added to these gene families or used to create new gene families with BLAST to determine homologous relationships ( $e\text{-value}>1e-10$  and  $\%id>50$ ). Multiple sequence alignments of extended gene families with *Lepidosteus oculatus* as an outgroup were produced using MAFFT<sup>62</sup> (command line option –auto) and gene trees were built with PhyML 3.4 (ref. 68) using the JTT+G substitution model. Using the NCBI teleost species tree, Softparsmap<sup>69</sup> was used to identify duplication and speciation event in trees. This resulted in 12,388 gene families with a speciation root node, encompassing 26,325 salmon genes.

The constructed gene trees were then assessed for duplicate retention for the Ts3R, Ss4R, small scale salmon specific duplications (SSD) following the Ss4R event, and duplications occurring between the Ts3R and Ss4R. Duplicate retention was counted by examining the conditional percentages of genes that were retained from the Ss4R following the Ts3R, and from the Ss4R to small-scale duplications on the salmon lineage. The duplication lineage for each gene was counted, ensuring that each lineage accounted for the retention or loss of a duplicate, with the expectation that each Ts3R duplication should give rise to two Ss4R, and every Ss4R should lead to two small scale duplications. Post3R–preSs4R SSDs also share an expectation of having resulted in two Ss4R duplications. Where nodes could be assigned as being either Ss4R or SSD, the chromosomal locations of the genes were used to differentiate between the ambiguous nodes. Such ambiguous nodes were determined to be SSDs if the duplicate salmon genes resided on the same chromosome; otherwise it was classified as being Ss4R. Since only a single Ss4R duplication occurred along a lineage, if two ambiguous nodes were found that could be classified as Ss4R along the same lineage, one was classified as being Ss4R

and the rest were classified as being SSD, with the oldest duplication being the Ss4R, an assumption that did not affect the trends in the data. Although most gene tree topologies were consistent with the teleost species tree, some gene trees showed large deviations from the accepted species tree. These trees may have been influenced by phylogenetic error which could cause spurious duplication counts and cause an overestimation of the number of duplication events within a gene family. Conditional probabilities were then calculated to determine the fraction of retained gene duplicates following each of the WGDs, given the opportunity for retention.

To assess if duplicate retention was impacted by protein–protein interactions, known protein–protein interactions were downloaded from the STRING database<sup>70</sup>. BLAST against *Danio rerio* was performed and putative STRING interactions in salmon were determined. Only interactions labelled ‘binding’ were kept, which are putative physical protein–protein interactions based on various forms of evidence. Patterns of co-retention following Ts3R, Ss4R, and SSD were then examined among STRING binding partners using the phylogenetic trees described above with custom perl scripts.

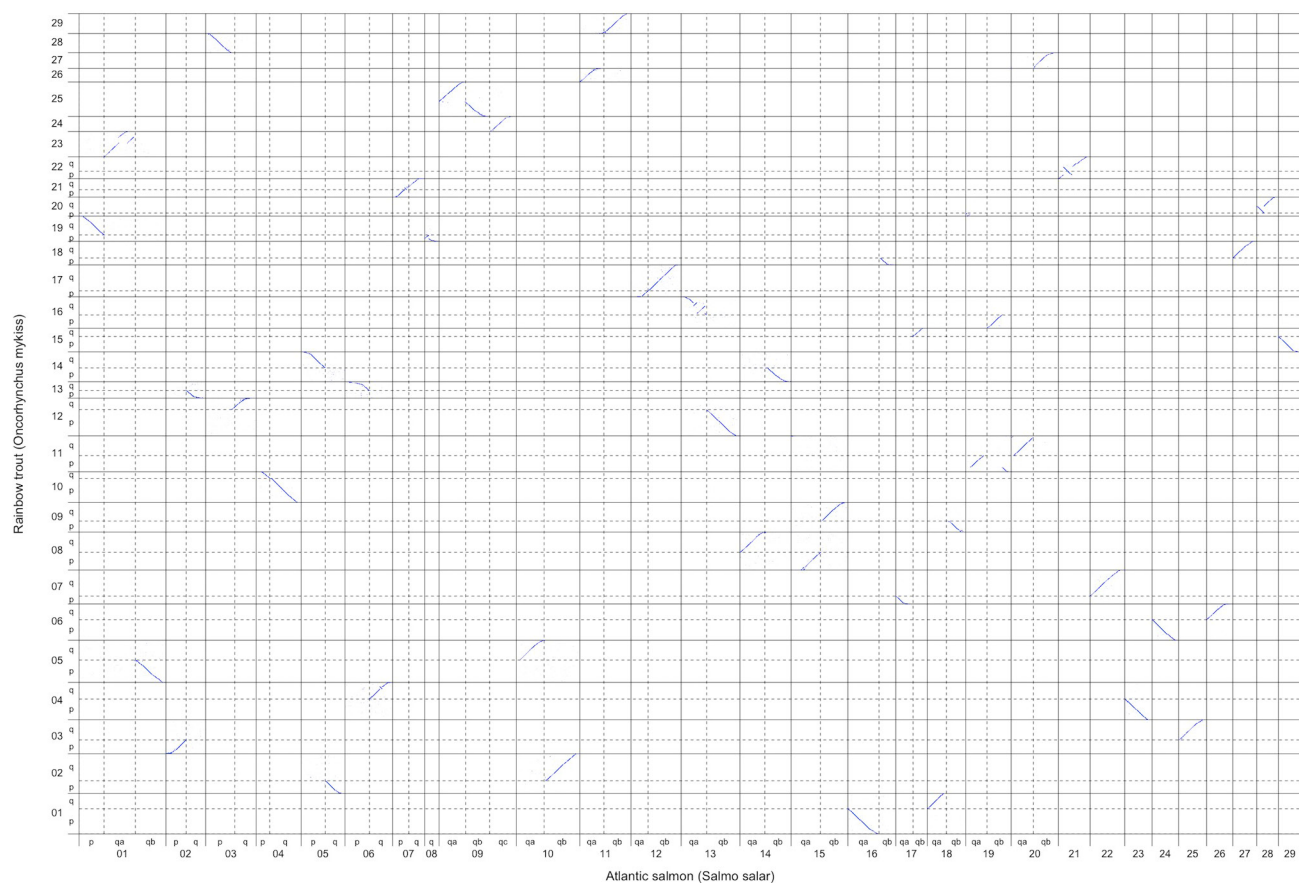
Statistical tests of significance were performed to determine if duplication counts were significantly different from each other. The duplication process was represented by a binomial distribution where each duplication could have either been retained or not. A two-proportion pooled z-test was performed to calculate two-sided *P* values at the Bonferroni corrected  $\alpha$ -level (0.001/7). To further explore if results were significant with a marginal effect level change or being overly influenced by large sample sizes, an odds ratio and relative risk analysis was performed for each group and two-sided *P* values were calculated. All tests showed extremely low *P* values indicating that the groups were significantly different from one another<sup>71</sup>. Effect sizes were considered as the fractional change in mean values.

All scripts used in this analysis are freely available on the Liberles Group website at Temple University (USA) at [https://liberles.cst.temple.edu/public/Salmon\\_Genome\\_Project/](https://liberles.cst.temple.edu/public/Salmon_Genome_Project/).

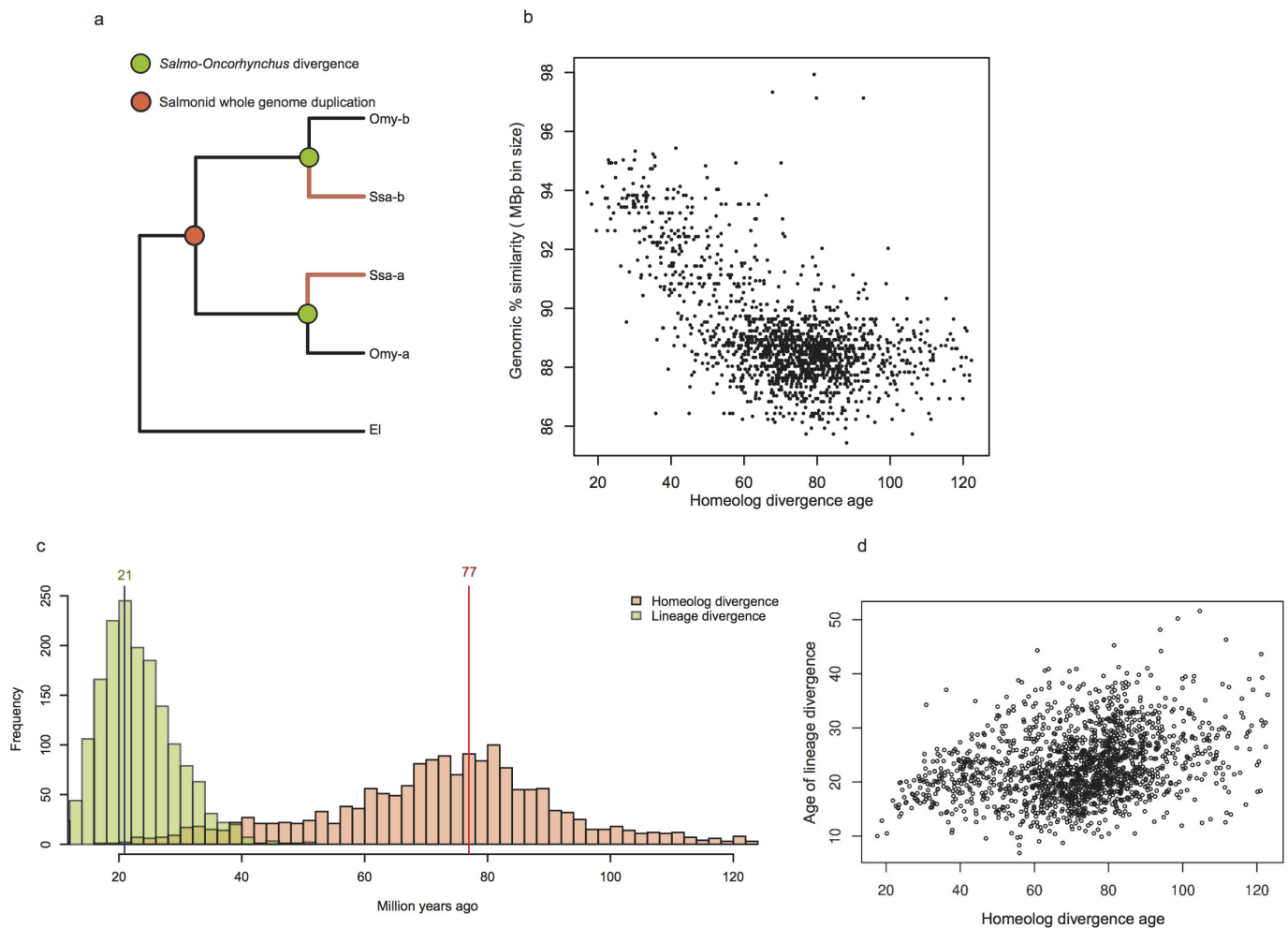
**Use of salmon assembly to improve rainbow trout genome sequence.** Salmon chromosome sequences were repeat masked using a salmon repeat database and RepeatMasker v4.0.3 (ref. 72) and aligned against rainbow trout scaffolds<sup>13</sup> using MegaBLAST<sup>73</sup>. Rainbow trout scaffolds mapping to multiple salmon chromosomes were broken when supported by information from a rainbow trout linkage map containing 31,390 SNPs constructed in a family material of 2,464 individuals using Lep-MAP<sup>74</sup>. The relative positions of trout scaffolds within the salmon genome were used, together with trout linkage maps, to position, orient and concatenate 11,335 rainbow trout scaffolds into 29 single chromosome sequences (1.37 Gb). Nomenclature for rainbow trout chromosomes is based on ref. 35. Conserved syntenic blocks between rainbow trout and Atlantic salmon were determined by aligning chromosome sequences for the two species against each other using LASTZ<sup>60</sup>.

37. Zimin, A. V. *et al.* The MaSuRCA genome assembler. *Bioinformatics* **29**, 2669–2677 (2013).
38. Green, P., Falls, K. & Crooks, S. *Documentation for CRI-MAP version 2.4*. (Washington University School of Medicine, 1990).
39. Leong, J. S. *et al.* *Salmo salar* and *Esoc lucius* full-length cDNA sequences reveal changes in evolutionary pressures on a post-tetraploidization genome. *BMC Genomics* **11**, 279 (2010).
40. Adzhubei, A. A. *et al.* Annotated expressed sequence tags (ESTs) from pre-smolt Atlantic salmon (*Salmo salar*) in a searchable data resource. *BMC Genomics* **8**, 209 (2007).
41. Koop, B. F. *et al.* A salmonid EST genomic study: genes, duplications, phylogeny and microarrays. *BMC Genomics* **9**, 545 (2008).
42. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
43. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
44. Wu, T. D. & Nacu, S. Fast and SNP-tolerant detection of complex variants and splicing in short reads. *Bioinformatics* **26**, 873–881 (2010).
45. Trapnell, C. *et al.* Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nature Protocols* **7**, 562–578 (2012).
46. Haas, B. J. *et al.* *De novo* transcript sequence reconstruction from RNA-seq: reference generation and analysis with Trinity. *Nature Protocols* **8**, 1494–1512 (2013).
47. Conesa, A. *et al.* Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics* **21**, 3674–3676 (2005).
48. Gremme, G., Brendel, V., Sparks, M. E. & Kurtz, S. Engineering a software tool for gene structure prediction in higher organisms. *Inf. Softw. Technol.* **47**, 965–978 (2005).
49. Jurka, J. *et al.* Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet. Genome Res.* **110**, 462–467 (2005).
50. Matveev, V. & Okada, N. Retroposons of salmonoid fishes (Actinopterygii: Salmonidae) and their evolution. *Gene* **434**, 16–28 (2009).
51. Ellinghaus, D., Kurtz, S. & Willhoeft, U. LTRharvest, an efficient and flexible software for *de novo* detection of LTR retrotransposons. *BMC Bioinformatics* **9**, 18 (2008).
52. Smit, A. F. A. & Hubley, R. *RepeatModeler Open-1.0*. <http://www.repeatmasker.org> (2008).
53. Flutre, T., Duprat, E., Feuillet, C. & Quesneville, H. Considering transposable element diversification in *de novo* annotation approaches. *PLoS ONE* **6**, e16526 (2011).
54. Camacho, C. *et al.* BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 421 (2009).
55. Jiang, N. Repeat Library Construction—Advanced <http://www.webcitation.org/6YWzgLCzw> (2013).
56. Wicker, T. *et al.* A unified classification system for eukaryotic transposable elements. *Nature Rev. Genet.* **8**, 973–982 (2007).
57. UniProt Consortium. UniProt: a hub for protein information. *Nucleic Acids Res.* **43**, D204–D212 (2015).
58. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32**, 1792–1797 (2004).
59. Pace, J. K., Gilbert, C., Clark, M. S. & Feschotte, C. Repeated horizontal transfer of a DNA transposon in mammals and other tetrapods. *Proc. Natl Acad. Sci. USA* **105**, 17023–17028 (2008).
60. Harris, R. S. *Improved Pairwise Alignment of Genomic DNA*. (ProQuest, 2007).
61. Krzywinski, M. *et al.* Circos: an information aesthetic for comparative genomics. *Genome Res.* **19**, 1639–1645 (2009).
62. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).
63. Schliep, K. P. phangorn: phylogenetic analysis in R. *Bioinformatics* **27**, 592–593 (2011).
64. Paradis, E., Claude, J. & Strimmer, K. APE: analyses of phylogenetics and evolution in R language. *Bioinformatics* (2004).
65. Yang, Z. PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* **24**, 1586–1591 (2007).
66. Drummond, A. J., Suchard, M. A., Xie, D. & Rambaut, A. Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Mol. Biol. Evol.* **29**, 1969–1973 (2012).
67. R Development Core Team. *R: A language and environment for statistical computing*. <https://www.r-project.org/> (R Foundation for Statistical Computing, 2009).
68. Guindon, S. *et al.* New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst. Biol.* **59**, 307–321 (2010).
69. Berglund-Sonnhammer, A.-C., Steffansson, P., Betts, M. J. & Liberles, D. A. Optimal gene trees from sequences and species trees using a soft interpretation of parsimony. *J. Mol. Evol.* **63**, 240–250 (2006).
70. Jensen, L. J. *et al.* STRING 8—a global view on proteins and their functional interactions in 630 organisms. *Nucleic Acids Res.* **37**, D412–D416 (2009).
71. Agresti, A. *Categorical Data Analysis* (John Wiley & Sons, 2002).
72. Smit, A. F. A., Hubley, R. & Green, P. RepeatMasker Open-4.0. <http://www.repeatmasker.org> (2013).
73. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).
74. Rastas, P., Paulin, L., Hanski, I., Lehtonen, R. & Auvinen, P. Lep-MAP: fast and accurate linkage map construction for large SNP datasets. *Bioinformatics* **29**, 3128–3134 (2013).





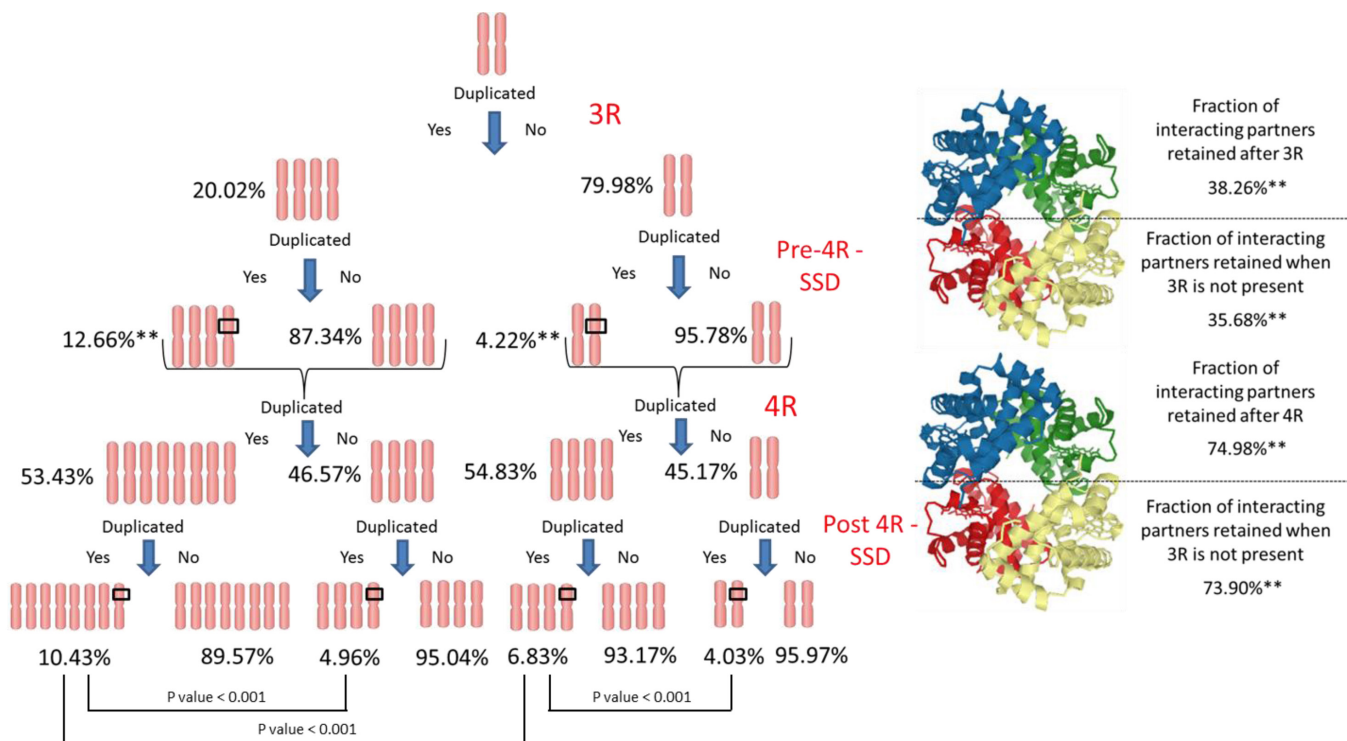
**Extended Data Figure 1 | Atlantic salmon and rainbow trout comparative map.** Alignment of Atlantic salmon (*Salmo salar*) and rainbow trout (*Oncorhynchus mykiss*) chromosome sequences using LASTZ demonstrates conservation of large collinear syntenic blocks between the two species.



**Extended Data Figure 2 | Dating or Ss4R rediploidization.** **a**, Schematic representation of a gene tree topology reflecting rediploidization of Ss4R homeologues before *Salmo-Oncorhynchus* divergence. **b**, Correlation between genomic similarity in 1 Mb windows and Ss4R rediploidization (that is, divergence) age. **c**, Distribution of *Salmo-Oncorhynchus*

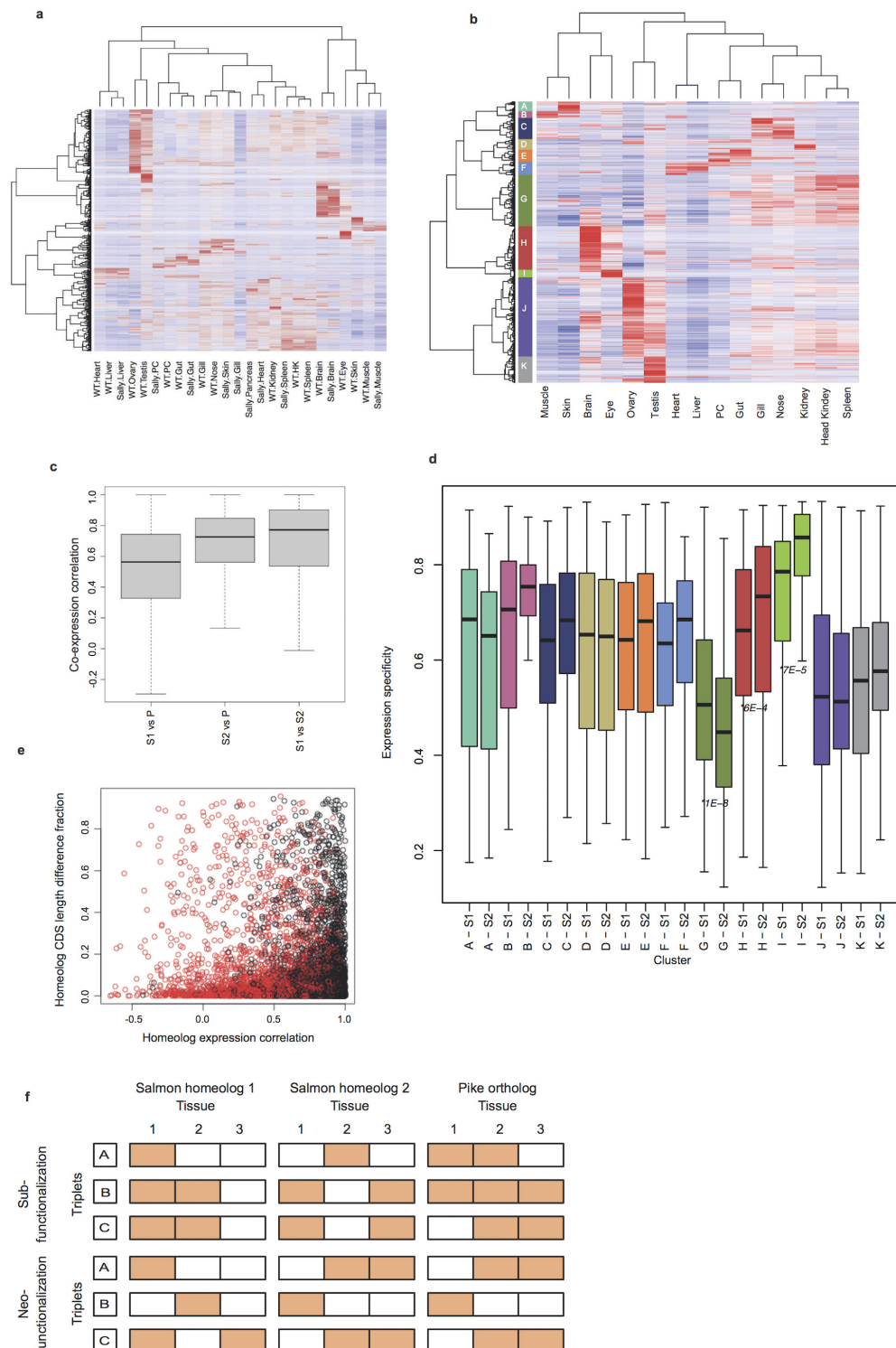
divergence age and Ss4R divergence age from time calibrated gene trees estimated with BEAST. Modes of each distribution are indicated with a vertical line. **d**, Correlation between estimated age of *Salmo-Oncorhynchus* divergence and Ss4R divergence age.





**Extended Data Figure 3 | Duplication count analysis and interacting partner co-retention.** The duplication process is depicted with the associated conditional probabilities for each type of duplication based upon a sampling of gene families that includes *Lepisosteus oculatus*. WGD events occur at both the Ts3R and Ss4R levels with individual gene duplications occurring at Pre-Ss4R-SSD and Post-Ss4R-SSD. Pre-Ss4R conditional probabilities are only dependent on Ts3R WGD being present and Ss4R WGD are only conditional on a Ts3R WGD being

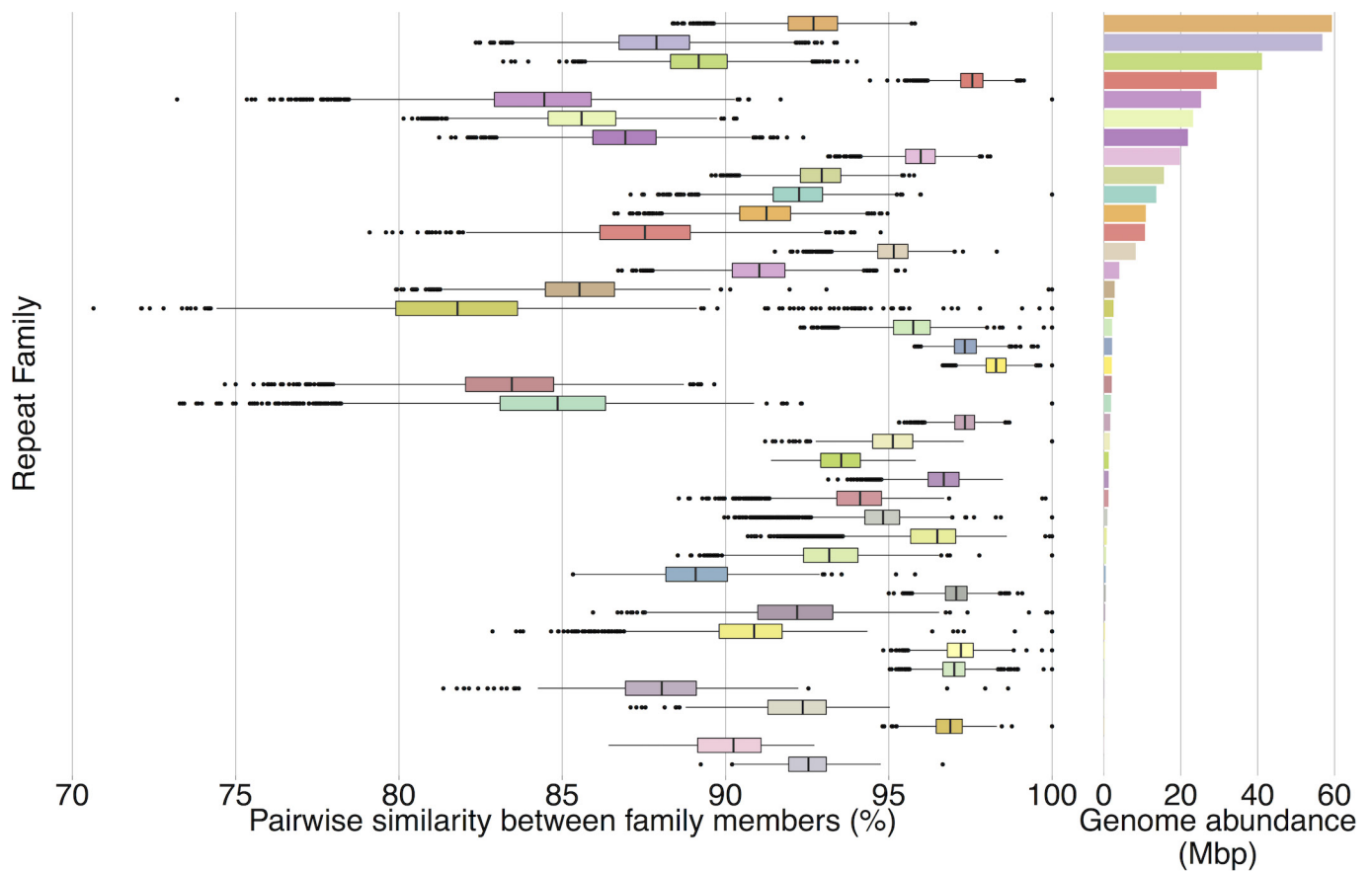
present. Retained interacting partners were determined from the STRING database<sup>48</sup> as partners with (binding) physical interaction. Interacting partners were determined based on being retained after the same Ts3R WGD or a Ss4R WGD as the query sequence and having a homologue in *Danio rerio*. Two asterisks indicate significance at  $\alpha < 0.001$  (Bonferroni corrected) based on a two-proportion pooled  $z$ -test from a binomial distribution.



#### Extended Data Figure 4 | Tissue gene expression regulation.

**a**, Hierarchical clustering of tissue gene expression in adult salmon from fresh water. WT = expression data from normal diploid Atlantic salmon. Sally = expression data from the double haploid fish used for reference genome sequencing. **b**, Classification of 11 co-expression clusters. Gene expression are from 15 tissues from a diploid adult Atlantic salmon from freshwater. Co-expression clusters are either associated with expression patterns from a single tissue or multiple tissues with similar physiological functions. Co-expression clusters A–K are named accordingly after the tissue(s) that contributes the most to its characteristic expression regulation profile: skin; skin and muscle; nose and gill; kidney; gut and pyloric ceca; heart and liver; unspecific; brain; eye; testis and ovary; testis. **c**, Gene expression correlation between salmon Ss4R homeologues and Northern pike orthologues. P = pike, S1 = salmon homeologue with lowest tissue expression correlation with

pike, S2 = salmon homeologue with highest tissue expression correlation to. **d**, Tissue expression specificity. Tissue expression specificity of Ss4R homeologues with novel gene regulation (S1) and conserved gene regulation (S2) compared to pike. Gene co-expression clusters are denoted A–K (see description in figure legend for b). Significantly different tissue specificity between diverged (S1) and conserved (S2) homeologues are indicated with a *P* value in the figure. **e**, Relationship between CDS-length difference and Ss4R expression regulation divergence. CDS length divergence are calculated as a fraction of the longest CDS in each Ss4R pair. Red colour represents homeologue pairs that are in different co-expression clusters (see above sections a and b for details). **f**, Illustration of sub- and neofunctionalization as defined by the analyses of ‘on’ and ‘off’ expression patterns. Red colour indicates a gene being ‘on’ in one tissue compared to its Ss4R duplicate and the assumed ancestral state of the diploid pike outgroup.



**Extended Data Figure 5 | Historical activity of 40 Tc1-mariner transposable elements and their abundance in the Atlantic salmon genome.**

Families with increased pairwise similarity between members have experienced less neutral sequence divergence since they were rendered inactive and reflect more recent additions to the genome.



# Sex-specific pruning of neuronal synapses in *Caenorhabditis elegans*

Meital Oren-Suisa<sup>1,2,3</sup>, Emily A. Bayer<sup>1,2,3</sup> & Oliver Hobert<sup>1,2,3</sup>

**Whether and how neurons that are present in both sexes of the same species can differentiate in a sexually dimorphic manner is not well understood. A comparison of the connectomes of the *Caenorhabditis elegans* hermaphrodite and male nervous systems reveals the existence of sexually dimorphic synaptic connections between neurons present in both sexes. Here we demonstrate sex-specific functions of these sex-shared neurons and show that many neurons initially form synapses in a hybrid manner in both the male and hermaphrodite pattern before sexual maturation. Sex-specific synapse pruning then results in the sex-specific maintenance of subsets of these connections. Reversal of the sexual identity of either the pre- or postsynaptic neuron alone transforms the patterns of synaptic connectivity to that of the opposite sex. A dimorphically expressed and phylogenetically conserved transcription factor is both necessary and sufficient to determine sex-specific connectivity patterns. Our studies reveal new insights into sex-specific circuit development.**

Like other invertebrate or vertebrate nervous systems, the nervous system of *C. elegans* contains a number of sex-specific neurons, most of which are generated during the process of sexual maturation in late larval stages<sup>1,2</sup>. Apart from these sex-specific neurons (8 in hermaphrodites, 91 in males), there are 294 neurons shared by both sexes, most of which are embryonically generated<sup>3,4</sup>. The hermaphrodite and male versions of these shared neurons display the same lineage history, the same cell body position, share molecular features (for example, neurotransmitter identity) and display similar neurite projection patterns<sup>3,5–7</sup>. Intriguingly, the recent reconstruction of the posterior nervous system of the *C. elegans* adult male<sup>5</sup> and its comparison to the connectome of the hermaphrodite (a derived female)<sup>8</sup> show that several of the sex-shared neurons are strongly sexually dimorphic in their synaptic wiring patterns. These anatomical observations provide a fascinating opportunity to study how seemingly similar sex-shared neurons develop sexually dimorphic characteristics.

## Synaptic target choices between sex-shared neurons

Here we focus on a group of sex-shared—but dimorphically connected—sensory, inter- and motorneurons<sup>5</sup> (Fig. 1a). The sexually dimorphic connectivity differences do not simply reflect sex-specific modifications of similar neuronal circuits, but rather sex-shared neurons wire into completely distinct circuits<sup>5</sup> (Fig. 1a). For example, the PHB phasmid sensory neuron synapses onto three different sex-shared command interneurons only in hermaphrodites (AVA, AVD and PVC). In males, PHB connects to a sex-shared interneuron, AVG, which in turn connects to downstream tail motor neurons only in males (Fig. 1a). To examine whether dimorphic connections are due to dimorphic synaptic partner choice or a reflection of sex-specific neuron or process placement, we analysed serial electron micrographs and found that the phasmid neuron processes are directly adjacent to the AVG process in both sexes (Extended Data Fig. 1). Dimorphic connections between these neurons are therefore a consequence of sex-specific synaptic partner choice.

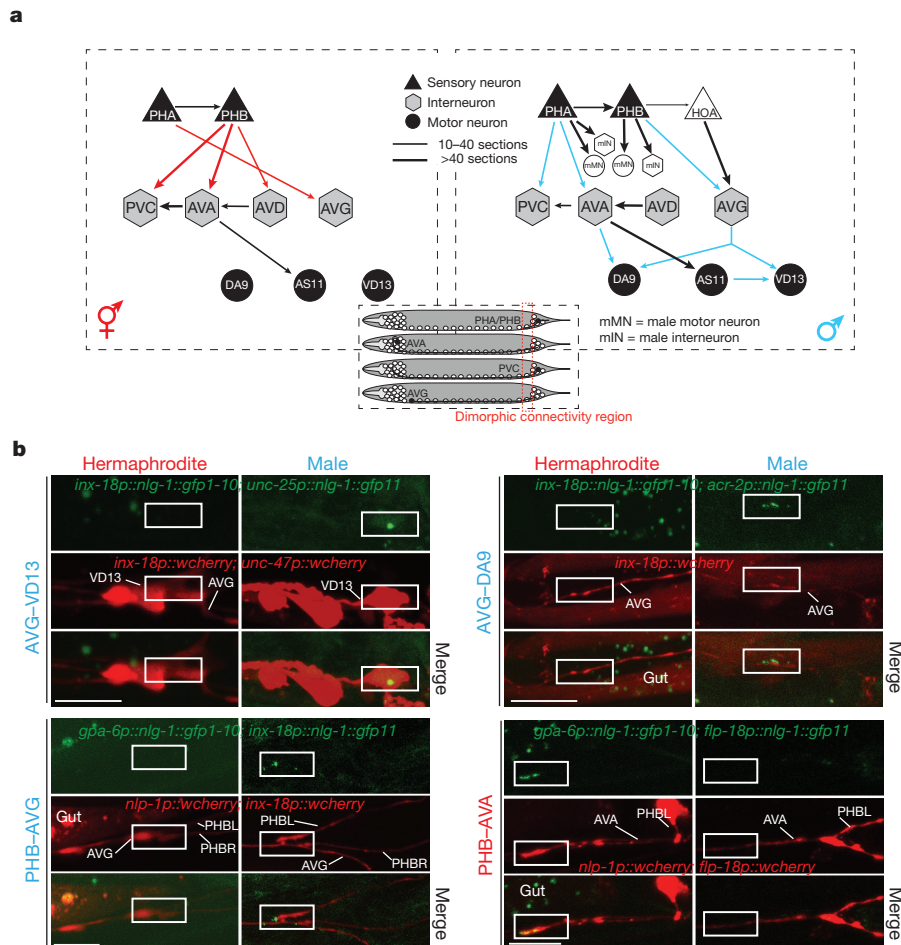
Although electron microscopy provides a powerful tool to identify synaptic partners, the presently available electron microscopy analysis relies on one or two animals at a single stage (adult). To confirm the electron microscopy results, and to visualize the reproducibility as

well as the developmental aspects of dimorphic synapses, we used two distinct trans-synaptic labelling techniques (Fig. 1b), ‘GRASP’ (GFP reconstitution across synaptic partners)<sup>9</sup> and ‘iBLINC’ (*in vivo* biotin labelling of intercellular contact)<sup>10</sup>. We generated transgenic lines in which seven distinct synaptically connected neuron pairs are labelled with GRASP and/or iBLINC. Using cytosolic mCherry to label neurites, we examined synaptic puncta along these mCherry-labelled adjacent processes. For all seven synaptic connections examined, we reproducibly identified discrete synaptic puncta that appear in the sexually dimorphic manner predicted by the electron microscopy analysis (Fig. 1a, b, Extended Data Figs 1–3).

## Sexually dimorphic functions of sex-shared neurons

To assess whether sexually dimorphic wiring is an indication of dimorphic neuronal function, we either surgically removed individual dimorphically connected neurons or genetically silenced them using ectopic expression of a histamine-gated chloride channel<sup>11</sup>. Silencing of the PHB neurons (using a driver that is strongly and consistently expressed only in PHB; Extended Data Fig. 4) affected forward locomotion of hermaphrodites, but not of males (Extended Data Fig. 5a). Another previously described PHB function also displays notable sex-specificity. Specifically, it has been shown that the PHB sensory neuron modulates chemorepulsive behaviour of hermaphrodites in response to the noxious chemical sodium dodecyl sulfate (SDS)<sup>12</sup>. The modulatory effect of PHB is observed upon functionally disabling subsets of head sensory neurons<sup>12</sup>. Such head-neuron-defective animals fail to avoid SDS because PHB provides an antagonistic input to command interneurons<sup>12</sup>. This antagonistic input to the command circuit is revealed through ablation of PHB, which restores the ability of head-sensory-neuron-defective animals to avoid SDS<sup>12</sup>. We corroborated this antagonistic input through silencing of PHB with a histamine-gated chloride channel, by examining *ceh-14* mutant animals, in which all three phasmid neurons fail to adopt their glutamatergic identity<sup>13</sup>, and by silencing both ASH and PHB neurons (Fig. 2a, Extended Data Fig. 5d, e). Importantly, male PHB does not connect to command interneurons and no antagonistic input to the command circuit is expected. One would therefore predict that the disabling of head sensory-neuron function in males (ASH silenced and

<sup>1</sup>Department of Biological Sciences, Columbia University, New York, New York 10027, USA. <sup>2</sup>Department of Biochemistry and Molecular Biophysics, Columbia University, New York, New York 10032, USA. <sup>3</sup>Howard Hughes Medical Institute, Columbia University, New York, New York 10027, USA.



**Figure 1 | Visualizing sexually dimorphic synapses. a**, Connectivity of selected neurons at the adult stage, as inferred from serial section reconstructions of electron micrographs<sup>5,8</sup>. Chemical synapses between sensory (triangles), inter- (hexagons) and motor (circles) neurons are depicted as arrows. Thickness of arrows correlates with degree of connectivity (number of sections over which *en passant* synapses are observed). The inset indicates where synaptic connections are formed. Black/grey, shared-sex neurons; white, male-specific neurons. Arrows indicate hermaphrodite-specific (red) and male-specific (blue) chemical synapse between shared neurons. **b**, Visualizing sexually dimorphic synapses. Fluorescent micrographs of GRASP GFP signal in preanal

ganglion region outlined in the inset in Fig. 1a. Neuronal processes are labelled with cytoplasmic codon-optimized Cherry markers. GRASP data are shown, additional iBLINC data are shown in Extended Data Fig. 3. Expression pattern of the promoters used in this study to drive cell-specific expression can be found in Extended Data Fig. 4. Quantification of data are shown in Fig. 3b; the number of fluorescent puncta (outlined with white boxes) is similar to those observed in the electron microscopy analysis<sup>5</sup>. Gut; auto-fluorescence gut granules. Scale bars, 10  $\mu$ m. In all images anterior is left, and dorsal is up. Blue and red indicate male and hermaphrodite, respectively, in all figures.

ablated; or *tax-4* mutants) should not have the impact on SDS avoidance that is observed in hermaphrodites. We indeed found this to be the case (Fig. 2a, Extended Data Fig. 5).

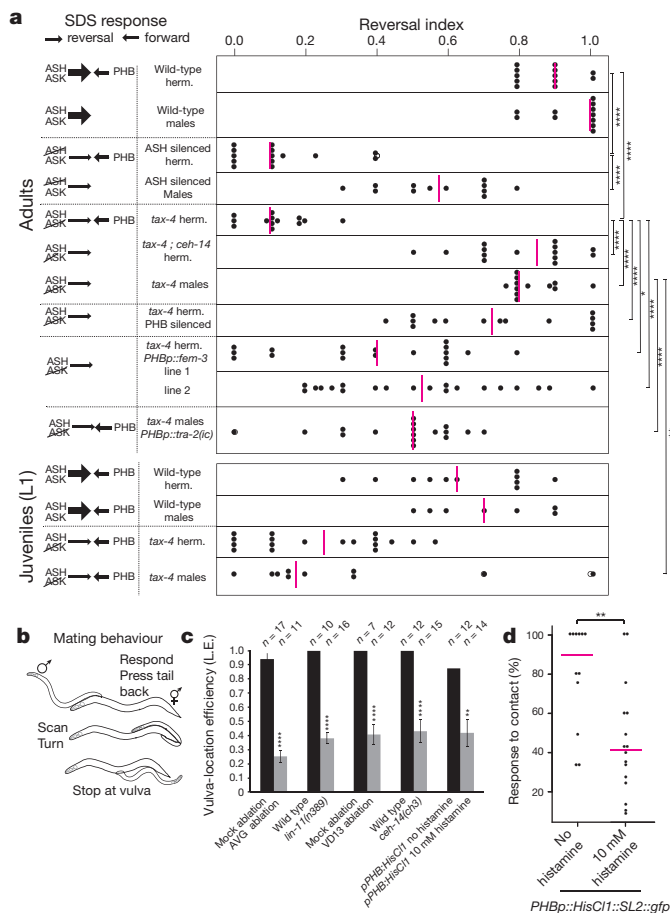
As adult male PHB neurons are not involved in the avoidance of noxious chemicals, we next investigated the function they adopt in males. We noted that the male-specific innervation target of PHB, the interneuron AVG, becomes innervated by male-specific hook sensory neurons (HOA and HOB), known to be involved in sensing a hermaphrodite-derived signal that induces males to stop at the vulva during mating behaviour<sup>14</sup>. Moreover, PHB also innervates HOA directly<sup>5</sup>. Through genetic silencing, genetic ablation and microsurgical laser ablations, we found that PHB and AVG are involved in responding to this vulva stop signal (Fig. 2b, c). Ablation of VD13, one of the motor neurons innervated by AVG specifically in males also caused vulva location defects (Fig. 2c), as did ablation of the LUA neuron (Extended Data Fig. 6). PHB also innervates the male-specific PVX and PVY neurons, involved in mate-contact-induced backward locomotion in males<sup>15</sup>, and silencing PHB also affected this sexually dimorphic behaviour (Fig. 2d). We conclude that PHB has sexually dimorphic functions, modulating locomotion and processing repulsive environmental sensory information in hermaphrodites, while

sensing hermaphrodite-derived mating cues in males (summarized in Extended Data Fig. 6c).

Sex-shared interneurons also display sexually dimorphic functions. We found that the dimorphically connected LUA neuron was also involved in response to vulva stop signal in males, and automated worm-tracking analysis<sup>16</sup> revealed that ablation of LUAs results in pausing defects in hermaphrodites but not in males (Extended Data Fig. 6).

### Sexually dimorphic synapse patterning

We next asked how sexually dimorphic connectivity patterns are established. With the exception of the VD13 neuron (which is born at the end of the first larval stage<sup>1</sup>), the shared neurons that we analysed are born and project their neurites during embryogenesis. Sex-specific neurons in the tail are born in the last larval stage<sup>2</sup> and male-specific behaviours emerge soon after<sup>17</sup>. One could therefore envision that dimorphic connections between the embryonically born neurons could form during sexual maturation, that is, long after the respective neurons and axonal patterns have been established in the animal ('late maturation' model; Fig. 3a). Alternatively, a 'default connectivity' could be established by both sexes early in development, followed by a sex-specific



**Figure 2 | Functional repurposing of dimorphic neurons.** **a**, Chemosensory repulsion assays (see Methods for full description of behavioural assays). Scatter diagrams plotting the avoidance index of single animals. Each black dot represents the fraction of reversal responses (scored as reversing or not reversing) in 10 or more assays of a single animal. Magenta vertical bars represent the median. The left column indicates predictions of reversal behaviour, based on previously published data that demonstrated a strong reversal drive from head neurons (thick arrows) that is counteracted by a forward drive mediated by the PHB neurons in the tail<sup>12</sup>. Control experiments for silencing using histamine and additional SDS assays can be found in Extended Data Fig. 5. *sra-6p::HisC11* and *gpa-6p::HisC11* were used for ASH and PHB silencing, respectively. Summary of dimorphic behaviours induced by PHB sensory neurons can be found in Extended Data Fig. 6c. **b**, Changes in male movement and posture triggered by mate contact<sup>14</sup>. **c**, Mutant or laser-operated animals tested for the male vulva location efficiency. In *lin-11* and *ceh-14* mutants, the AVG and phasmid neurons, respectively, fail to differentiate<sup>13,23</sup>. Error bars, s.e.m. **d**, Initiation of backward movement in response to hermaphrodite contact is dependent on PHB activity, measured as contact response efficiency. Each dot represents one animal. We performed the nonparametric Mann–Whitney test (Wilcoxon rank sum test) with Bonferroni correction for multiple comparisons (**a**, **c**) and Fisher’s exact test (**d**). \*\*\*\* $P < 0.0001$ , \*\*\* $P < 0.001$ , \*\* $P < 0.01$ , \* $P < 0.05$ ; NS, not significant.

synapse rewiring upon sexual maturation (‘rewiring’ model; Fig. 3a). By analysing connectivity throughout larval development using the GRASP and iBLINC systems, we discovered two different mechanisms: a ‘pruning’ mechanism, used by many different neurons, and a ‘prepat- terning’ mechanism used by others (Fig. 3a).

For example, PHB neurons connect to AVA and AVG in hermaphrodites and males at early larval stages (Fig. 3b, c). Therefore, male-specific pruning of the PHB–AVA synapses results in an adult male-specific PHB–AVG connection, and conversely, hermaphrodite-specific pruning of the PHB–AVG synapses results in

the hermaphrodite-specific PHB–AVA synapses. We observed these synaptic pruning events to occur during sexual maturation in the fourth larval stage (L4) (Extended Data Fig. 7).

As the PHB–AVA connection exists in both sexes at the L1 stage, we asked whether at this stage both hermaphrodites and males display a PHB-dependent modulation of the repulsive response to noxious chemicals. We indeed found this to be the case (Fig. 2a). The repulsive response was restored in both sexes upon silencing of PHB (Extended Data Fig. 5f). These observations demonstrate that PHB does not merely acquire dimorphic functions in the adult, but that PHB neurons in males undergo a repurposing of function, from initially being involved in sensing noxious environmental cues to processing sex-specific cues.

Sexually dimorphic pruning of synaptic connections can also be observed in a number of additional neuronal contexts (Fig. 3, Extended Data Fig. 2). For example, PHA connects to AVG in both sexes at the L1 stage, but this connection is selectively lost in males (Fig. 3b, c). Similarly, the AVG interneuron connects to the cholinergic DA9 motor neuron in both sexes at the L1 stage, but the synaptic connection persists only in males, not hermaphrodites (Figs 1b, 3b, c).

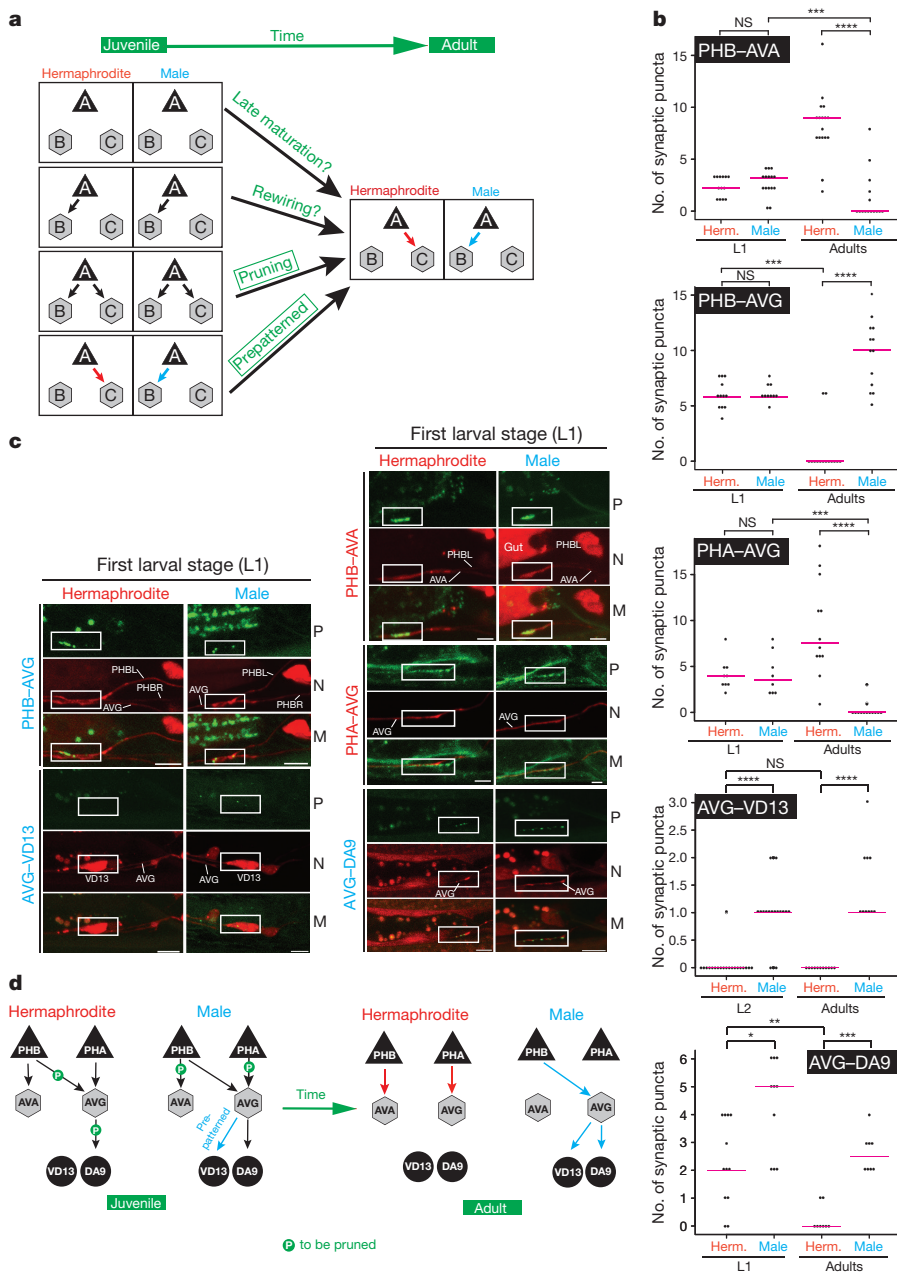
One other sexually dimorphic synaptic connection arises by a fundamentally different principle: the AVG to VD13 connection is only ever observed in males, and never in hermaphrodites (Figs 1b and 3b, c). Some prepat- terning is also already evident in the pruned AVG–DA9 synapses; these are present in both sexes in the L1 stage, but are stronger in the male (Fig. 3b, c). Analysis of ~1,000 serial electron microscopy sections shows that the axons of the AVG and VD13 (synaptically connected in a male-specific manner) are more adjacent in males compared to hermaphrodites (Extended Data Fig. 1). Even though we cannot exclude the possibility that dimorphic adjacency is merely a secondary consequence of failure to establish synaptic contact, we propose that the prepat- terning of dimorphic synapse may be a consequence of dimorphic axon placement. Taken together, we identified two types of synaptic maturation events: pruning events that coincide with sexual maturation and dimorphic prepat- terning events that precede sexual maturation (summarized in Fig. 3d).

## The sex of neurons controls synaptic patterning

We next investigated whether sexually dimorphic wiring patterns are determined by the sex of both the pre- and postsynaptic cell, or by non-cell-autonomous processes. We addressed this question by generating sexually mosaic animals through cell-type specific, ectopic expression of the *fem-3* gene, which downregulates the global hermaphroditic identity determinant TRA-1 (a Gli-type zinc finger transcription factor) thereby imposing a male identity on the specific cell-type in an otherwise hermaphroditic animal<sup>6,18,19</sup>. Conversely, ectopic cell-type-specific expression of the intracellular domain of the TRA-2 receptor (‘TRA-2<sup>IC</sup>’) feminizes cells via stabilization of the TRA-1 transcription factor in otherwise male animals<sup>17,20</sup>. We found that identity transformations of single neurons transformed synaptic wiring patterns to that of the opposite sex (Fig. 4, Extended Data Fig. 8). For example, masculinization of PHB results in a loss of the PHB–AVA connection in an otherwise hermaphroditic animal and a gain of the PHB–AVG connection, demonstrating that the sex of PHB dictates which connection is pruned or maintained. Non-dimorphic PHB connectivity at the L1 stage is not affected by sex-reversal (Extended Data Fig. 8). Notably, masculinization of PHB also restores the behavioural defects of head-sensory-neuron-disabled hermaphrodites (Fig. 2a). Thus, the behavioural differences between males and hermaphrodites in the noxious chemical response can be linked specifically to the sex of an individual neuron (Fig. 2).

Not only is the sex of the presynaptic neuron a determinant of sex-specific wiring patterns, but the postsynaptic cell is also important for establishing proper connectivity (Fig. 4). Masculinization of AVG in hermaphrodites (in which PHB does not normally connect





**Figure 3 | Synaptic patterning during development.** **a**, Models of how sexually dimorphic connectivity patterns may arise during development. **b**, **c**, Quantification and fluorescent micrographs of synaptic pruning measured by the number of synaptic puncta observed using GRASP GFP (PHB-AVA, PHB-AVG, AVG-VD13, AVG-DA9) and iBLINC GFP (PHA-AVG) in L1 and adult hermaphrodites and in males. As VD neurons are born at the end of the L1 stage, juvenile VD13 puncta were quantified at the L2 stage. We performed nonparametric Mann-Whitney test (Wilcoxon rank sum test) with Bonferroni correction for multiple comparisons. \*\*\*\* $P < 0.0001$ , \*\*\* $P < 0.001$ , \*\* $P < 0.01$ , \* $P < 0.05$ ; NS, not significant. Magenta horizontal bars represent the median. Region of neurite overlap and observed synaptic puncta marked with white boxes. Gut, auto-fluorescence gut granules. Scale bars, 5 μm. Note that the roughly twofold increase in synapse number from L1 to L4 in the hermaphroditic PHB-AVA connection and the male PHB-AVG is in line with an overall increase in total synapse numbers seen between neurons between L1 and adult stage as deduced by recent reconstruction of an L1 stage animal (M. Zhen, personal communication). **d**, Summary of synaptic connection differences between juvenile, pre-L4 animals and adults.

to AVG) maintains the PHB-AVG connection. Similarly, masculinization of AVA disrupts maintenance of the PHB-AVA synapses in hermaphrodites (Fig. 4).

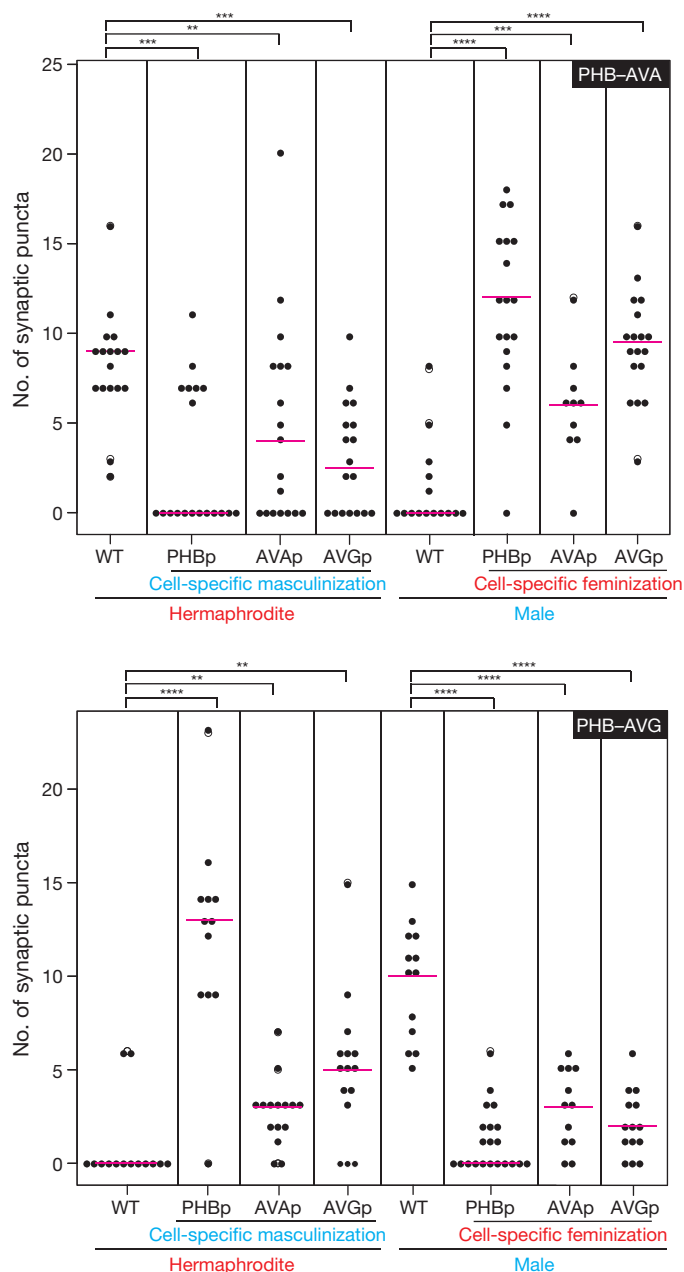
Feminizing neurons cell-autonomously via expression of TRA-2<sup>IC</sup> in otherwise male animals generally produced the converse effects (Fig. 4). For example, feminization of the PHB neurons in male animals is sufficient to maintain the normally hermaphrodite-specific PHB-AVA synapse and prune the normally male-specific PHB-AVG synapses (Fig. 4). These ectopic PHB-AVA synapses seem to be functional, as *tax-4* males with feminized PHBs reverse less and can antagonize the SDS response (Fig. 2a).

We further probed the non-autonomous nature of maintaining sex-specific synapses by asking whether the maintenance of the PHB-AVG synapses in hermaphrodites, through masculinization of AVG, affects the maintenance of the hermaphrodite-specific PHB-AVA synapse. Indeed, in animals in which *fem-3* is driven in AVG (resulting in stabilization of the PHB-AVG synapse), PHB-AVA synapse number is significantly reduced. Conversely, in hermaphrodites in which we masculinized AVA, the PHB-AVA synapses are not only pruned, but the PHB-AVG synapses are stabilized (Fig. 4). These results suggest a

competition mechanism in which one synaptic wiring configuration is maintained at the expense of the 'alternative' wiring pattern.

### Doublesex-like transcription factors control dimorphic connectivity

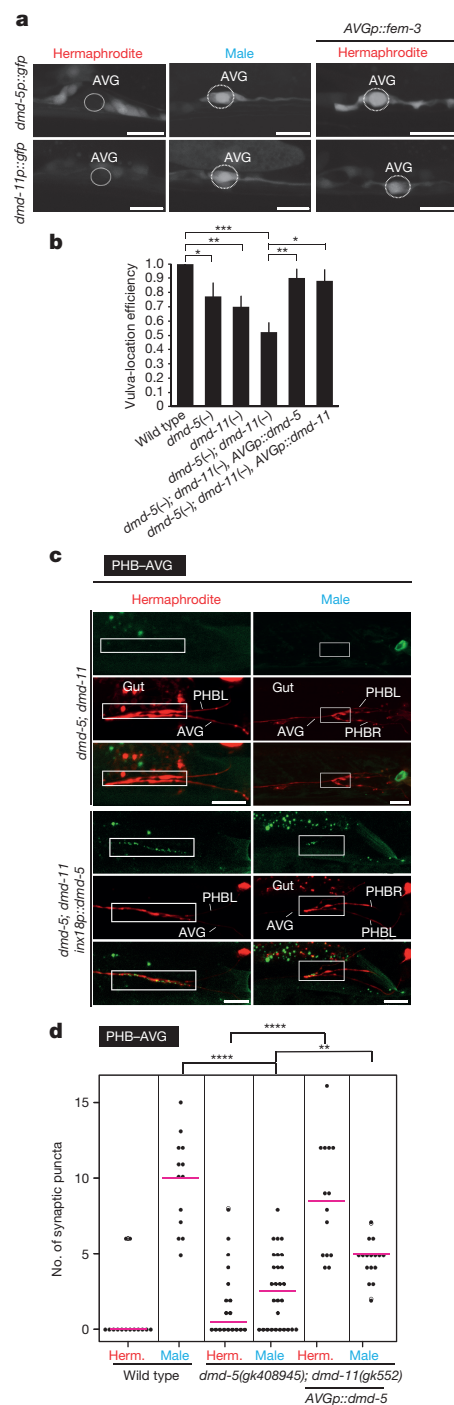
We next sought to determine the link between the globally acting sex determination system, mediated by TRA-1, and synaptic pruning. A recently framed hypothesis posits that globally acting hormonal signals in vertebrates operate through region-specific modular effector system<sup>21</sup>. In *C. elegans*, the global TRA-1 regulator may operate in a cell-type-specific manner through one of the 11 members of the DMD (Doublesex/MAB-3 domain) family of transcription factors, which are conserved regulators of sexual identity in various organisms<sup>22</sup>. Three of these family members (*mab-3*, *mab-23*, *dmd-3*) have previously been implicated in somatic sex differences<sup>6</sup>, but the other eight have remained uncharacterized. We found that *dmd-5* and *dmd-11* are dimorphically expressed in the dimorphically connected AVG neuron described above; expression is observed in male AVG, but not in hermaphrodite AVG (Fig. 5a). This dimorphic expression is controlled cell-autonomously via the canonical sex-determination



**Figure 4 | Autonomy and non-autonomy of sex-specific synapse pruning.** Either the presynaptic cell (PHB) or the postsynaptic cells (AVA, AVG) were masculinized (by expression of TRA-2<sup>IC</sup>) and feminized (by expression of FEM-3), and the number of synaptic puncta were quantified in hermaphrodites and males. We performed nonparametric Mann–Whitney test (Wilcoxon rank sum test) with Bonferroni correction for multiple comparisons. \*\*\*\* $P < 0.0001$ , \*\*\* $P < 0.001$ , \*\* $P < 0.01$ . Magenta horizontal bars represent the median. Results are summarized in Extended Data Fig. 8a. Note that a general trend in the sex-reversal experiments is that the change of the sex of the presynaptic neuron (PHB) appears to have a stronger effect than changing the sex of either of the postsynaptic cells. Changing the sex of both postsynaptic cells simultaneously did not enhance the effects (Extended Data Fig. 8c).

pathway and TRA-1, as expression of FEM-3—the negative regulator of hermaphroditic TRA-1 protein—in the AVG neurons derepresses *dmd-5* and *dmd-11* expression in hermaphrodites (Fig. 5a, Extended Data Fig. 9a).

*dmd-5* and *dmd-11* single mutants show defects in the male-specific function of AVG in mating behaviour (Fig. 5b, Extended Data Fig. 9f–h) and *dmd-5*; *dmd-11* double mutant animals show even



**Figure 5 | Sexually dimorphic expression and function of *dmd-5* and *dmd-11*.**

**a**, Sex-specific expression of *dmd-5* and *dmd-11* reporter genes in AVG. There are no additional dimorphisms in the retrovascular ganglion, apparent differences are due to differences in z-planes incorporated into the final Z-stack projection. Masculinization of AVG derepresses *dmd-5* and *dmd-11* expression in hermaphrodites AVG. Quantified in Extended Data Fig. 9a. **b**, Vulva location efficiency is affected in *dmd-5* ( $n = 11$ ), *dmd-11* ( $n = 20$ ) and *dmd-5*; *dmd-11* double mutant males ( $n = 29$ ) compared with wild-type males ( $n = 15$ ). Expression of either *dmd-5* ( $n = 10$ ) or *dmd-11* ( $n = 10$ ) in AVG (using the *inx-18* promoter) of double mutant males rescues behaviour defects. Error bars, s.e.m. **c**, **d**, *dmd-5* and *dmd-11* are required for maintenance of AVG synapses. Fluorescent micrographs (c) and quantification of synaptic puncta of PHB–AVG (d). Region of neurite overlap and observed synaptic puncta marked with white boxes. Quantification of DMD effect on LUA–AVG synapses can be found in Extended Data Fig. 9i. We performed nonparametric Mann–Whitney test (Wilcoxon rank sum test) with Bonferroni correction for multiple comparisons (b, d). \*\*\*\* $P < 0.0001$ , \*\*\* $P < 0.001$ , \*\* $P < 0.01$ , \* $P < 0.05$ . Magenta horizontal bars represent the median. Scale bars, 10  $\mu$ m.

stronger defects (Fig. 5b). These defects can be rescued by AVG-specific expression of either *dmd-5* or *dmd-11* (Fig. 5b). Moreover, *dmd-5* single mutants and *dmd-5; dmd-11* double mutants display the alterations in AVG synaptic wiring that one would expect from factors that control the sexually dimorphic nature of AVG wiring (Fig. 5c, d, Extended Data Fig. 9). Male-specific PHB–AVG synapses fail to be maintained in *dmd-5; dmd-11* mutant males. Synaptic defects can be rescued through AVG-specific expression of *dmd-5* (Fig. 5c, d). There are no synaptic defects observed in the L1 stage of *dmd-5; dmd-11* mutants when no synaptic dimorphism is yet apparent (Extended Data Fig. 9b), indicating that *dmd-5* and *dmd-11* are not required for synapse formation *per se*, but are specifically involved in controlling sex-specific synapse maintenance by preventing synaptic pruning, as predicted by their expression pattern. Furthermore, the PHB–AVA hermaphroditic connection is non-autonomously stabilized in *dmd-5; dmd-11* mutant males, supporting the competition model discussed above (Extended Data Fig. 9e).

*dmd-5* is not only required but also sufficient to prevent synaptic pruning, as deduced by ectopic *dmd-5* expression in the AVG neurons of hermaphrodites. In these animals the PHB–AVG synapses are maintained (Fig. 5d). As DMD proteins are generally thought to work as repressors<sup>22</sup>, we propose that *dmd-5*, in conjunction with *dmd-11*, represses the expression of gene(s) in male AVG neurons that are involved in the pruning of AVG connections to PHB and that repression of these pruning factor(s) in hermaphroditic AVG, via ectopic *dmd-5* expression, inhibits pruning (summarized in Extended Data Fig. 9j). Notably, although AVG-masculinized *dmd-5*(+) hermaphrodites (masculinized through *fem-3*-driven degradation of TRA-1) do maintain PHB–AVG synapses, AVG-masculinized *dmd-5*(−) hermaphrodites do not (Fig. 5d, Extended Data Fig. 9d). This demonstrates that *dmd-5* functions downstream of *tra-1*, as already suggested by the observation of ectopic *dmd-5* expression in animals in which we degraded TRA-1 cell-autonomously in AVG (see above; Fig. 5a, Extended Data Fig. 9a).

In conclusion, our studies show how sex-shared neurons adopt sex-specific synaptic wiring patterns. Sex-specific wiring patterns arise in a neuron-type specific manner. In most cases, we observe sexual maturation-coupled sex-specific pruning of synaptic connections that were indiscriminately generated in both sexes. In at least one case, sex-specific synapses are prepatterned before overt sexual maturation of the animal. We observed notable patterns of cellular autonomy of the synaptic pruning events and we found pruning to be regulated by sex-specifically expressed transcription factors.

**Online Content** Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

**Received 7 August 2015; accepted 6 April 2016.**

**Published online 4 May 2016.**

1. Sulston, J. E. & Horvitz, H. R. Post-embryonic cell lineages of the nematode, *Caenorhabditis elegans*. *Dev. Biol.* **56**, 110–156 (1977).
2. Sulston, J. E., Albertson, D. G. & Thomson, J. N. The *Caenorhabditis elegans* male: postembryonic development of nongonadal structures. *Dev. Biol.* **78**, 542–576 (1980).
3. Sulston, J. E., Schierenberg, E., White, J. G. & Thomson, J. N. The embryonic cell lineage of the nematode *Caenorhabditis elegans*. *Dev. Biol.* **100**, 64–119 (1983).
4. Sammut, M. *et al.* Glia-derived neurons are required for sex-specific learning in *C. elegans*. *Nature* **526**, 385–390 (2015).

5. Jarrell, T. A. *et al.* The connectome of a decision-making neural network. *Science* **337**, 437–444 (2012).
6. Emmons, S. W. The development of sexual dimorphism: studies of the *Caenorhabditis elegans* male. *Wiley Interdiscip. Rev. Dev. Biol.* **3**, 239–262 (2014).
7. Portman, D. S. Genetic control of sex differences in *C. elegans* neurobiology and behavior. *Adv. Genet.* **59**, 1–37 (2007).
8. White, J. G., Southgate, E., Thomson, J. N. & Brenner, S. The structure of the nervous system of the nematode *Caenorhabditis elegans*. *Phil. Trans. R. Soc. Lond. B* **314**, 1–340 (1986).
9. Feinberg, E. H. *et al.* GFP Reconstitution Across Synaptic Partners (GRASP) defines cell contacts and synapses in living nervous systems. *Neuron* **57**, 353–363 (2008).
10. Desbois, M., Cook, S. J., Emmons, S. W. & Bulow, H. E. Directional *trans*-synaptic labeling of specific neuronal connections in live animals. *Genetics* **197**, 697–705 (2015).
11. Pokala, N., Liu, Q., Gordus, A. & Bargmann, C. I. Inducible and titratable silencing of *Caenorhabditis elegans* neurons *in vivo* with histamine-gated chloride channels. *Proc. Natl Acad. Sci. USA* **111**, 2770–2775 (2014).
12. Hilliard, M. A., Bargmann, C. I. & Bazzicalupo, P. *C. elegans* responds to chemical repellents by integrating sensory inputs from the head and the tail. *Curr. Biol.* **12**, 730–734 (2002).
13. Serrano-Saiz, E. *et al.* Modular control of glutamatergic neuronal identity in *C. elegans* by distinct homeodomain proteins. *Cell* **155**, 659–673 (2013).
14. Liu, K. S. & Sternberg, P. W. Sensory regulation of male mating behavior in *Caenorhabditis elegans*. *Neuron* **14**, 79–89 (1995).
15. Sherlekar, A. L. *et al.* The *C. elegans* male exercises directional control during mating through cholinergic regulation of sex-shared command interneurons. *PLoS ONE* **8**, e60597 (2013).
16. Yemini, E., Jucikas, T., Grundy, L. J., Brown, A. E. & Schafer, W. R. A database of *Caenorhabditis elegans* behavioral phenotypes. *Nature Methods* **10**, 877–879 (2013).
17. Mowrey, W. R., Bennett, J. R. & Portman, D. S. Distributed effects of biological sex define sex-typical motor behavior in *Caenorhabditis elegans*. *J. Neurosci.* **34**, 1579–1591 (2014).
18. White, J. Q. *et al.* The sensory circuitry for sexual attraction in *C. elegans* males. *Curr. Biol.* **17**, 1847–1857 (2007).
19. Lee, K. & Portman, D. S. Neural sex modifies the function of a *C. elegans* sensory circuit. *Curr. Biol.* **17**, 1858–1863 (2007).
20. Mehra, A., Gaudet, J., Heck, L., Kuwabara, P. E. & Spence, A. M. Negative regulation of male development in *Caenorhabditis elegans* by a protein–protein interaction between TRA-2A and FEM-3. *Genes Dev.* **13**, 1453–1463 (1999).
21. Yang, C. F. & Shah, N. M. Representing sex in the brain, one module at a time. *Neuron* **82**, 261–278 (2014).
22. Matson, C. K. & Zarkower, D. Sex and the singular DM domain: insights into sexual regulation, evolution and plasticity. *Nature Rev. Genet.* **13**, 163–174 (2012).
23. Hutter, H. Extracellular cues and pioneers act together to guide axons in the ventral cord of *C. elegans*. *Development* **130**, 5307–5318 (2003).

**Supplementary Information** is available in the online version of the paper.

**Acknowledgements** We thank Q. Chen for generating transgenic strains; J. White, J. Sulston and LMB/MRC for sharing their annotated electron microscopy images to D. Hall for curation, and <http://www.wormimage.org> where these annotated images have been made available by D. Hall; E. Yemini for advice on tracking experiments; S. Cook for help with Elegance software; M. Zhen for communicating unpublished results; P. Sengupta, I. Greenwald and members of the Hobert lab for comments on the manuscript. This work was supported by postdoctoral fellowships from the EMBO and HFSP (to M.O.-S.), the NIH (2R37NS039996) and the HHMI. M.O.-S. is an Awardee of the Weizmann Institute of Science, National Postdoctoral Award Program for Advancing Women in Science.

**Author Contributions** M.O.-S. and O.H. designed the experiments. M.O.-S. performed most experiments. E.A.B. quantified the data for PHA-AVG synapses and all iBLINC transgenes, tracked silenced PHB animals and generated driver lines for expression analysis of *gpa-6* and *flp-18*. M.O.-S. and O.H. wrote the paper with input from E.A.B.

**Author Information** Reprints and permissions information is available at [www.nature.com/reprints](http://www.nature.com/reprints). The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to M.O.-S. (meitals@gmail.com) or O.H. (or38@columbia.edu).



## METHODS

**Data reporting.** No statistical methods were used to predetermine sample size. The experiments were not randomized and the investigators were not blinded to allocation during outcome assessment.

**C. elegans strains.** Wild-type strains were *C. elegans* variety Bristol, strain N2. Worms were maintained according to standard methods<sup>24</sup>. Worms were grown at 20°C on nematode growth media (NGM) plates seeded with bacteria (*E. coli* OP50) as a food source. Mutant strains used in this study include: CB1489 *him-8(e1489)* IV, CB4088 *him-5(e1490)* V, MT633 *lin-11(n389)* I; *him-5(e1467)* V, RB1295 *F10C1.5(ok1394)/+* II, FX01760 *dmd-5(tm1760)/+* II, PR678 *tax-4(p678)* III, DA509 *unc-31(e928)* IV, TB528 *ceh-14(ch3)* X, VC30074 *dmd-5(gk408945)* II, VC1193 *dmd-11(gk552)* V.

All transgenic strains used in this study are listed in Supplementary Table 1, ordered by figures and extended data figures.

**Cloning and constructs.** To generate *inx-18p::wCherry* (pMO10), *inx-18* second intron was amplified from pOH260 (ref. 25) and restriction sites were added to ends (5'-SphI, 3'-PstI). This fragment was digested and ligated into pPD95.75 vector in which the GFP was replaced with codon-optimized mCherry ('worm Cherry'). To restrict expression to AVG, AIY motif (5'-ATTAGTTTCGTTAA-3') was deleted from the 2<sup>nd</sup> intron of *inx-18* using site directed mutagenesis (Forward primer: 5'-AATTTTTCATGTTACCTACTTATTTTCTATTAGCGTCATAG AT-3', Reverse: 5'-ATCTATGACGCTAATAGAAAAAATAAGTAGGTAACATG AAAAAAATT-3'). *inx-18* 2<sup>nd</sup> intron is also dimly and variably expressed in URX.

*fem-3::SL2* was amplified from *P(rab-3)::fem-3::SL2::mCherry* Gateway construct<sup>18</sup> using PCR and KpnI restriction site was added to 5' end. This PCR was digested and inserted into KpnI; AscI digested pPD95.75 with wCherry, to generate *fem-3::SL2::wCherry* (pMO11).

To generate *inx-18p::fem-3::SL2::wCherry* (pMO12), pMO10 was digested using SphI, PstI and ligated into digested pMO11.

To generate *inx-18p::nlg-1::gfp1-10* (pMO17), *inx-18p::nlg-1::gfp11* (pMO18), *inx-18p::BirA::nrx-1* (pMO19) and *inx-18p::AP::nlg-1* (pMO24), *inx-18p* was cloned from pMO10 (*inx-18p::wCherry*) into the SphI, XmaI sites of MVC2 (pSM::nlg-1::GFP1-10), MVC3 (pSM::nlg-1::gfp11) (a gift from M. VanHoven), *gcy-8::BirA::nrx-1* and *ttx-3p::AP::nlg-1* (a gift from H. Bulow), respectively.

*eat-4p9* ('LUA promoter', a fragment of 172 bp upstream to the *eat-4* gene) was amplified using PCR from *eat-4p2* (ref. 11) and restriction sites were added to ends (5'-HindIII, 3'-BamHI). This fragment was digested and ligated into pPD95.75 vector, to generate *eat-4p9::gfp* (pMO13). *eat-4p9* is expressed in the tail neurons LUAs and PVR.

To generate *eat-4p9::nlg-1::gfp1-10* (pMO14), *eat-4p9* was amplified from *eat-4p2* and restriction sites were added to ends (5' XbaI, 3' NheI). This fragment was digested and ligated into MVC2.

To generate *eat-4p9::fem-3::SL2::wCherry* (pMO15), pMO14 was digested using SpeI, NheI and the fragment containing *eat-4p9* was ligated into pMO11.

*eat-4p9::wCherry* was generated by PCR fusion of *eat-4p9* and wCherry<sup>26</sup>.

To generate *eat-4p9::BirA::nrx-1* (pMO16), *eat-4p9* was amplified from *eat-4p2* and restriction sites were added (5'-XmaI, 3'-SphI). This fragment was digested and ligated into digested *gcy-8::BirA::nrx-1*.

To generate *gpa-6p::nlg-1::gfp11* (pMO21), *gpa-6p::BirA::nrx-1* (pMO22) and *gpa-6p::fem-3::SL2::wCherry* (pMO29), 2.6 kb *gpa-6* promoter was cloned from MVC6 (pSM::gpa-6p::nlg-1::gfp1-10, a gift from M. VanHoven) into the SphI, SmaI sites of MVC3 (pSM::nlg-1::GFP11), *gcy-8p::BirA::nrx1* and pMO11, respectively. To generate *gpa-6p::gfp* the 2.6 kb *gpa-6* fragment was cloned into the SphI, SmaI sites of pPD95.75. *gpa-6* is expressed brightly and consistently in PHBs and also dimly and variably in AWAs.

To generate *flp-18p::fem-3::SL2::2XNLS::TagRFP* (pMO30), FEM-3 was cloned into an *SL2::2XNLS::TagRFP* plasmid using restriction-free (RF) cloning<sup>27</sup>. The 3.1 kb *flp-18* PCR product was cloned into the SphI, XmaI sites of *fem-3::SL2::2XNLS::TagRFP*. To generate *flp-18p::gfp*, the 3.1 kb *flp-18* PCR product was cloned into the SphI, XmaI sites of pPD95.75. *flp-18* is expressed brightly and consistently in AVAs, and dimly and variably in AIYs.

To generate *srg-13p::BirA::nrx-1* (pMO23), a 3 kb promoter fragment of *srg-13* was amplified from genomic DNA (primer F; 5'-GTACCTGCAGAA GGACTTGG CAGAAAGAAGC-3', R: 5'-TGACCCGGGTGGGCTGTAATTT TTAGCTCG-3'), with PstI, XmaI sites added to ends. This was cloned into the PstI, XmaI sites of pPD95.75. A 2.2 kb *srg-13* promoter was then cloned into the SphI, XmaI sites of *gcy-8p::BirA::nrx-1*.

To generate *acr-2p::AP::nlg-1* (pMO25), *acr-2p* was cloned from pEVL194 (*acr-2p::nlg-1::gfp11*, a gift from B. D. Ackley) into the SphI, XmaI sites of *ttxp-3::AP::nlg-1*.

To generate the cell-specific feminizing constructs, *tra-2(ic)::SL2::2XNLS::TagRFP* (pMO37) was made by inserting *tra-2(ic)* from pENTRY 1-2 *tra-2(ic)*

(a gift from D. Portman) into an *SL2::2XNLS::TagRFP* construct using RF cloning. *inx-18p::tra-2(ic)::SL2::2XNLS::TagRFP* (pMO32) was generated by subcloning *inx-18p* from pMO10 (*inx-18p::wCherry*) into the SphI, XmaI sites of pMO37. *gpa-6p::tra-2(ic)::SL2::2XNLS::TagRFP* (pMO33) was generated by subcloning *gpa-6* promoter from MVC6 into the SphI, SmaI sites of pMO37. *flp-18p::tra-2(ic)::SL2::2XNLS::TagRFP* (pMO34) was generated by subcloning 3.1 kb *flp-18* PCR product from MVC 12 into the SphI, XmaI sites of pMO37. *eat-4p9::tra-2(ic)::SL2::2XNLS::TagRFP* was generated by ligating the digested *eat-4p9* mentioned above into the SphI, XmaI sites of pMO37.

To generate PHB histamine-induced silencing construct, a 2.6 kb *gpa-6* promoter from MVC6 was cloned into pNP403 (*tag-168p::HisCl1::SL2::gfp*; a gift from N. Pokala and C. Bargmann) to replace the *tag-168* promoter using RF cloning.

To generate *dmd-5* and *dmd-11* genomic rescue constructs, *dmd-5* and *dmd-11* genomic sequences including 500 bp 3'-UTR were amplified from genomic DNA and cloned into pMO10 to replace wCherry using RF cloning. The resulting constructs are pMO31, *inx-18p::dmd-5 genomic+3'-UTR* and pMO38, *inx-18p::dmd-11 genomic+3'-UTR*.

**Microscopy.** Worms were anaesthetized using 100 mM of sodium azide (NaN<sub>3</sub>) and mounted on 5% agar on glass slides. Worms were analysed by Nomarski optics and fluorescence microscopy, using a Zeiss 780 confocal laser-scanning microscope. When using GFP, we estimated the resolution of our confocal to be ~250 nm. Multidimensional data were reconstructed as maximum intensity projections using Zeiss Zen software. Puncta were quantified by scanning the original full Z-stack for distinct dots in the area where the processes of the two neurons overlap. Figures were prepared using Adobe Photoshop CS6 and Adobe Illustrator CS6.

**Cell ablation.** We performed laser ablations using a MicroPoint Laser System Basic Unit (N2 pulsed laser (dye pump), ANDOR Technology) attached to a Zeiss AxioPlan 2IE widefield microscope (objective EC Plan-Neofluar 100×/1.30 Oil M27). This laser delivers 120 μJoules of 337 nm energy with a 3-nsec pulse length. Ablations were performed as previously described<sup>28</sup>, with pulse repetition rates of ~15 Hz. Cell identification was performed with GFP or Cherry markers. Ablations were performed at the L4 stage, and worms were analysed 24–48 h later. Mock animals were placed on same slide under microscope but were not ablated, and were allowed to recover in a similar manner. After relevant assays were performed (tracking or mating assays), worms were mounted again on glass slides and analysed under microscope to validate that cell-ablation was successful.

**Mating behaviour assays.** Mating assays were based on procedures described previously<sup>12,29</sup>. Males were picked at an early L4 stage and kept apart from hermaphrodites for 24 h. One male was transferred to a plate covered with a thin fresh OP50 lawn containing 10–15 adult *unc-31(e928)* hermaphrodites. These hermaphrodites move very little, allowing for an easy recording of male behaviour. Hermaphrodites were also isolated from opposite sex at the L4 stage and used 24 h later. Animals were monitored and sequence of events was recorded within a 15 min window or until the male ejaculated, whichever occurred first. Males were tested for their ability to locate vulva in a mating assay, calculated as location efficiency<sup>30</sup>. The number of passes or hesitations at the vulva until the male first stops at the vulva were counted: location efficiency = 1/number of encounters to stop. PHB-silenced males were digitally recorded using the Exo Labs model 1 camera mounted on Nikon Eclipse E400 compound microscope with long-distance X20 lens. These videos were analysed for vulva location efficiency and percentage of successful contact responses, which requires tail apposition and initiation of backward locomotion. Percentage response to contact = 100 × (the number of times a male exhibited contact response/the number of times the male makes contact with a hermaphrodite via the rays)<sup>15</sup>.

**SDS-avoidance behaviour.** SDS-avoidance assay was based on procedures described previously<sup>12</sup>. A small drop of solution containing either the repellent (0.1% SDS in M13 buffer) or buffer (M13 buffer: 30 mM Tris-HCl (pH 7.0), 100 mM NaCl, 10 mM KCl) is delivered near the tail of an animal while it moves forward. Once in contact with the tail, the drop surrounds the entire animal by capillary action and reaches the anterior amphid sensory organs. The drop was delivered using 10 μl glass calibrated pipets (VWR international) pulled by hand on a flame to reduce the diameter of the tip. The capillary pipette was mounted in a holder with rubber tubing and operated by mouth. Assayed worms were transferred to fresh non-wet unseeded NGM plates and allowed to rest at room temperature for a few minutes. Each assay started with testing the animals with drops of buffer alone. The response to each drop was scored as reversing or not reversing. The avoidance index is the number of reversal responses divided by the total number of trials. An inter-stimuli interval of at least two minutes was used between successive drops to the same animal. Each animal was tested at least 10 times. The L1 animals were scored blindly and sex was noted 48 h later.

**Neuronal silencing using the histamine chloride channel 1 (HisCl1) system.** *ASH::HisCl1* (*kyEx5104*), *PHBp::HisCl1* (*otEx6341*), *PHBp::HisCl1*; *tax-4(p678)*

and *ASH::HisCl1*; *PHBp::HisCl1* transgenic animals were picked at the L4 stage and placed on NGM plates containing 10 mM histamine with OP50 bacteria as food source. As a control, animals were placed on NGM plates containing OP50 bacteria but no histamine. Chemorepulsion behaviour assays were performed 24 h later. Histamine plates were prepared as previously described<sup>31</sup>. As additional controls, panneuronal *HisCl1* transgenic worms (*CX14373 kyEx4571 (pNP403 (tag-168::HisCl1::SL2::gfp), myo-3::mCherry)*; a gift from C. Bargmann) were placed on histamine plates prior to use. Histamine plates on which *tag-168::HisCl1* animals were paralyzed under a minute were used for behavioral assays. There was no difference in the avoidance index of wild-type worms grown on histamine plates and on plates without histamine.

**Automatic worm tracking.** Hermaphrodite and male transgenic *otIs462* animals were ablated at the L4 stage and left to recover for 24 h before tracking. PHB-silenced hermaphrodites and males were transferred into NGM plates containing 10 mM histamine for 24 h before tracking. At the adult stage, animals were placed on an NGM plate seeded with diluted 20 µl of OP50 bacteria in the centre. As a control, mock ablated transgenic *otIs462* animals were used. Automatic tracking was performed at ~22°C (room temperature) with Worm Tracker 2.0 (WT2), which uses a mobile camera to track and record individual worms and 5 min videos were generated. Analysis of the tracking videos was performed as previously described<sup>16</sup>.

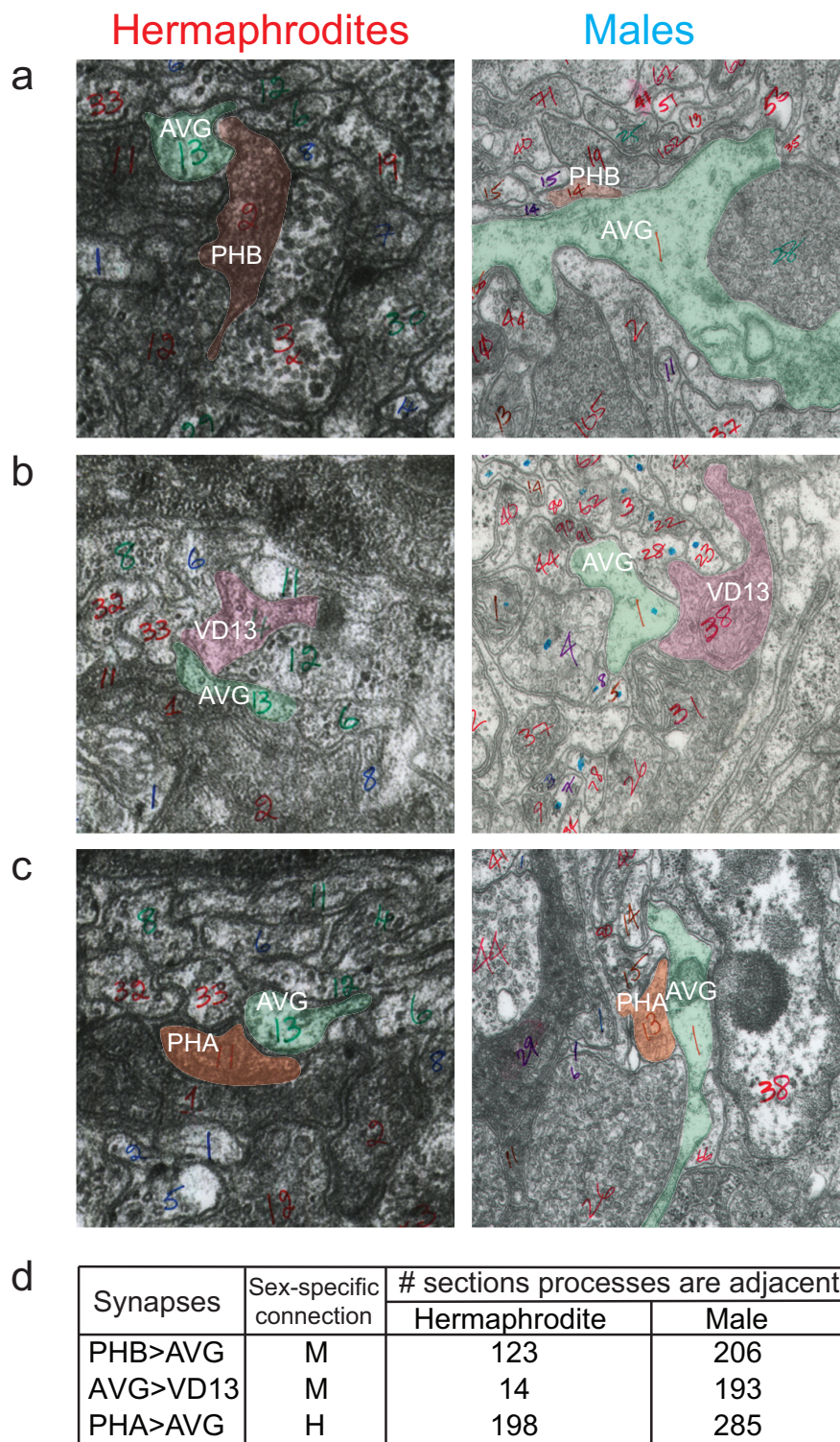
To assess the effect of LUA ablation, we first conducted a pilot tracking study of LUA-ablated hermaphrodites worms versus controls and LUA-ablated males versus controls. Owing to the extensive analysis of the tracking system, 702 features were initially measured. After correction for multiple testing, pausing and pausing-related features emerged as the main, significant difference between LUA-ablated males and hermaphrodites. Therefore, to obviate further heavy corrections for multiple testing, we chose to re-run these experiments, this time only measuring pausing, permitting us to use of the uncorrected *P* value. Within this subset of new experiments, wherein only one feature was measured (that is, pausing), we found highly significant *P* values. Given the necessity of four tests to make the proof, we chose the most conservative correction possible for multiple testing, Bonferroni, to illustrate that the *P* values measured maintain significance even after such extreme correction.

**Analysis of electron microscopy data.** Original transmission electron microscopy serial electron micrographs from 'JSE' hermaphrodite tail series and 'N2Y' male tail series were analysed using Elegance software<sup>32</sup>.

Images were aligned and screened for adjacency of neuronal processes of the pairs of cells described in this manuscript. For JSE, images from 'PAG' and 'red series' were analysed, and for N2Y, 'PAG' images were analysed.

24. Brenner, S. The genetics of *Caenorhabditis elegans*. *Genetics* **77**, 71–94 (1974).
25. Wenick, A. S. & Hobert, O. Genomic *cis*-regulatory architecture and *trans*-acting regulators of a single interneuron-specific gene battery in *C. elegans*. *Dev. Cell* **6**, 757–770 (2004).
26. Hobert, O. PCR fusion-based approach to create reporter gene constructs for expression analysis in transgenic *C. elegans*. *Biotechniques* **32**, 728–730 (2002).
27. Bond, S. R. & Naus, C. C. RF-Cloning.org: an online tool for the design of restriction-free cloning projects. *Nucleic Acids Res.* **40**, W209–W213 (2012).
28. Fang-Yen, C., Gabel, C. V., Samuel, A. D., Bargmann, C. I. & Avery, L. Laser microsurgery in *Caenorhabditis elegans*. *Methods Cell Biol.* **107**, 177–206 (2012).
29. Garcia, L. R., LeBoeuf, B. & Koo, P. Diversity in mating behavior of hermaphroditic and male–female *Caenorhabditis nematodes*. *Genetics* **175**, 1761–1771 (2007).
30. Peden, E. M. & Barr, M. M. The KLP-6 kinesin is required for male mating behaviors and polycystin localization in *Caenorhabditis elegans*. *Curr. Biol.* **15**, 394–404 (2005).
31. Gordus, A., Pokala, N., Levy, S., Flavell, S. W. & Bargmann, C. I. Feedback from network states generates variability in a probabilistic olfactory circuit. *Cell* **161**, 215–227 (2015).
32. Xu, M. *et al.* Computer assisted assembly of connectomes from electron micrographs: application to *Caenorhabditis elegans*. *PLoS ONE* **8**, e54050 (2013).
33. White, G. Neuronal connectivity in *Caenorhabditis elegans*. *Trends Neurosci.* **8**, 277–283 (1985).
34. Durbin, R. M. Studies on the development and organisation of the nervous system of *Caenorhabditis elegans*. PhD thesis, University of Cambridge (1987).
35. Rogers, C. *et al.* Inhibition of *Caenorhabditis elegans* social feeding by FMRFamide-related peptide activation of NPR-1. *Nature Neurosci.* **6**, 1178–1185 (2003).
36. Komatsu, H., Mori, I., Rhee, J. S., Akaike, N. & Ohshima, Y. Mutations in a cyclic nucleotide-gated channel lead to abnormal thermosensation and chemosensation in *C. elegans*. *Neuron* **17**, 707–718 (1996).
37. Thompson, O. *et al.* The million mutation project: a new approach to genetics in *Caenorhabditis elegans*. *Genome Res.* **23**, 1749–1762 (2013).
38. Murphy, M. W. *et al.* An ancient protein-DNA interaction underlying metazoan sex determination. *Nature Struct. Mol. Biol.* **22**, 442–451 (2015).



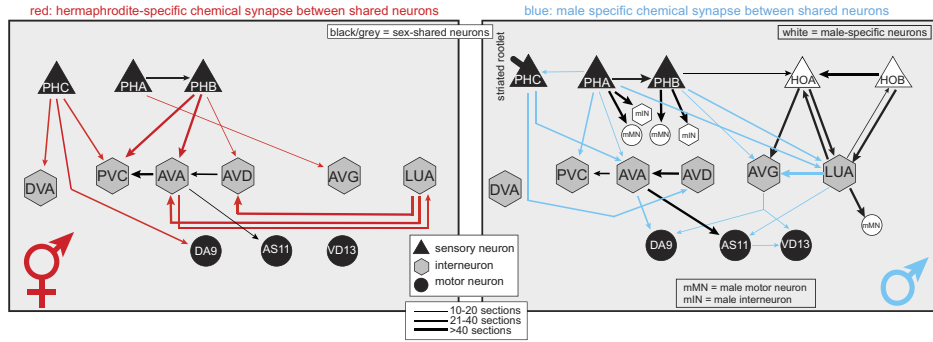


**Extended Data Figure 1 | Adjacency of neuronal processes in hermaphrodites and males.** Four transmission electron microscopy prints from wild-type adult hermaphrodite 'JSE'<sup>8</sup> and four from adult male 'N2Y', showing adjacency of neuronal processes. These images were collected at MRC/LMB for ref. 8 and ref. 2, and the annotated images are available online at <http://www.wormimage.org>, courtesy of D. Hall. The set of processes directly adjacent to one another has been defined as the 'neighbourhood' of that process<sup>33</sup>, and the placement of processes into specific neighbourhoods is a major determinant of connectivity. Connections form only in one sex, although processes are adjacent in both sexes. **a**, Print 385, JSE series (JSE\_122283; <http://wormimage.org/image.php?i=122283&page=2>) and print 620, N2Y series (PAG620; <http://wormimage.org/image.php?id=103528&page=18>) shows

PHB–AVG adjacent processes, pseudo labelled in green (AVG) and red (PHB). **b**, Print 359, JSE series (JSE\_122257; <http://wormimage.org/image.php?id=122257&page=2>) and print 500, N2Y series (PAG500; <http://wormimage.org/image.php?id=103408&page=20>) shows AVG–VD13 adjacent processes, pseudo labelled in green (AVG) and pink (VD13). **c**, Print 377, JSE series (JSE\_122275; <http://wormimage.org/image.php?id=122275&page=2>) and print 800, N2Y series (PAG800; <http://wormimage.org/image.php?id=103706&page=16>) shows PHA–AVG adjacent processes, pseudo labelled in green (AVG) and orange (PHA). **d**, A table summarizing the number of electron microscopy sections in which direct adjacency of processes was observed. Over a 1000 PAG (preanal ganglion) serial sections were analysed for each sex.



a

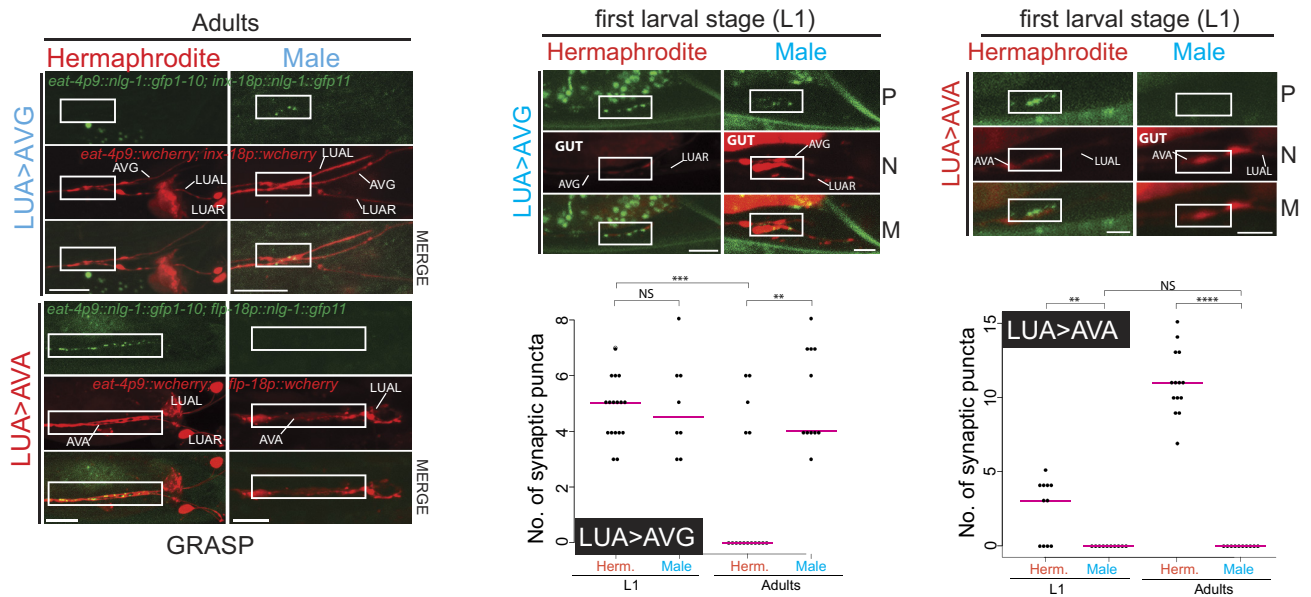


**b**

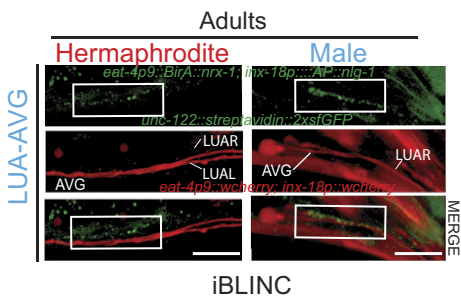
GRASP and/or iBLINC

Pre	Post	Sex	Presynaptic marker	Postsynaptic marker
LUA	AVG	♂	<i>eat-4p9::nlg-1::spGFP1-10</i> <i>eat-4p9::BirA::nrx-1</i>	<i>inx-18p::nlg-1::spGFP11</i> <i>inx-18p::AP::nlg-1</i>
	AVA	♀	<i>eat-4p9::nlg-1::spGFP1-10</i>	<i>flp-18p::nlg-1::spGFP11</i>

C

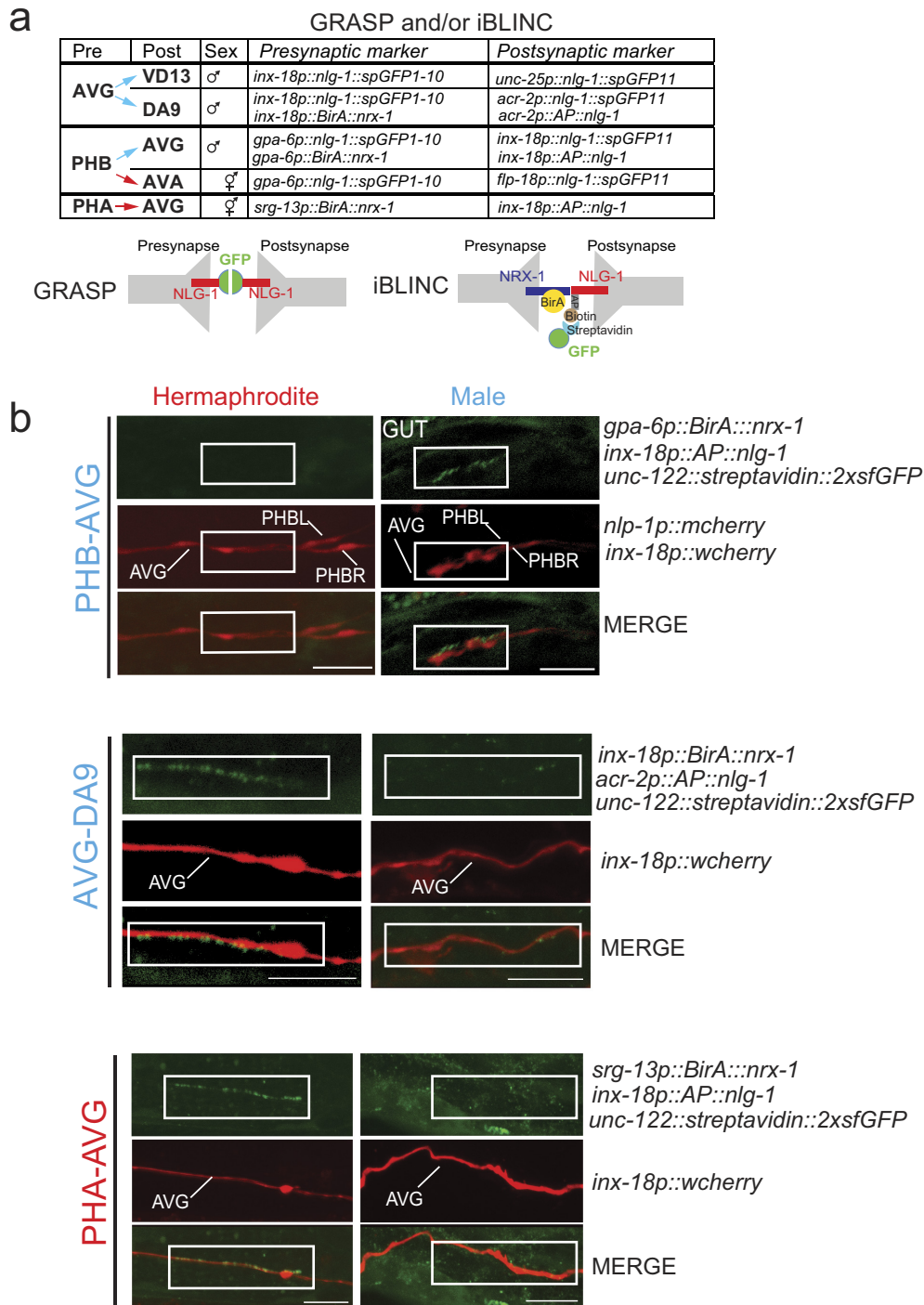


d



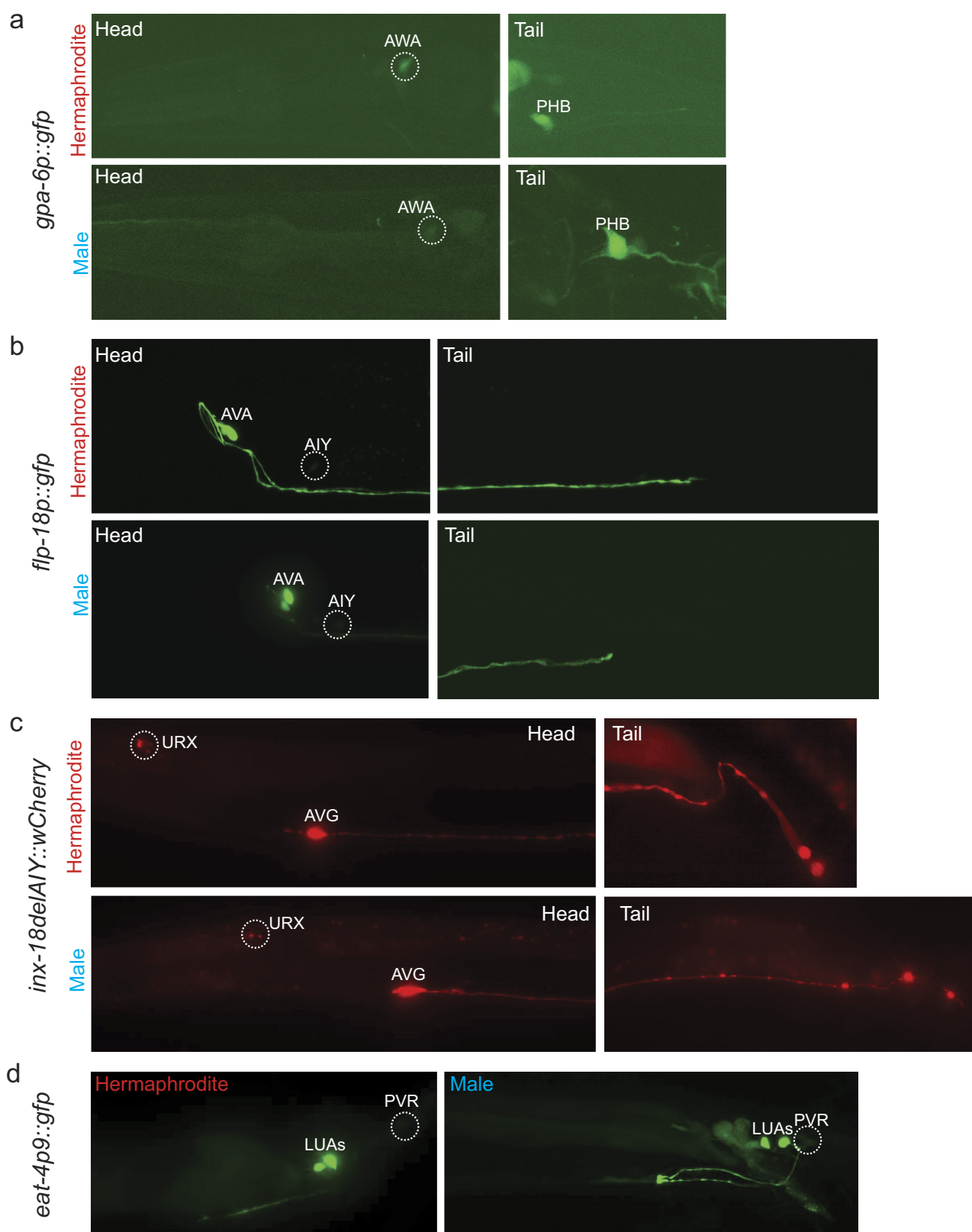
**Extended Data Figure 2 | Dimorphic connections of LUA.** **a**, The connectivity diagram shown in Fig. 1a, including dimorphic connections of the LUA and PHC connections. The GRASP data that we show in this figure as well as the pruning data, sexual reversal data and mutant data shown in Extended Figs 6, 8 and 9, supports the original LUA connectivity data reported in ref. 5 and summarized in this schematic. However, a recent reassessment of the tracing of electron micrographs suggests that the connectivity assignments of the PHC and LUA neurons may have been swapped with each other (S. Emmons, personal communication). **b**, Overview of LUA synaptic connections labelled in this paper. **c**, Visualizing LUA sexually dimorphic synapses. Quantification and fluorescent micrographs of GRASP trans-synaptically labelled puncta

between LUA–AVG and LUA–AVA, in L1 and adult hermaphrodites and in males. M, merge; N, neurite; P, puncta. Region of neurite overlap and observed synaptic puncta are marked with white boxes. Gut, auto-fluorescence gut granules. Scale bars, 10  $\mu$ m (adult) and 5  $\mu$ m (L1). **d**, Fluorescent micrographs of the preanal ganglion region of transgenic animals expressing the presynaptic *BirA::nrx-1* fusion in LUA (using the *eat-4p9* promoter), and postsynaptic receptor *peptide::nlg-1* fusion in AVG. For more details see Extended Data Fig. 3. Scale bars, 10  $\mu$ m. We performed nonparametric Mann–Whitney test (Wilcoxon rank sum test) with Bonferroni correction for multiple comparisons. \*\*\*\* $P < 0.0001$ , \*\*\* $P < 0.001$ , \*\* $P < 0.01$ ; NS, not significant. Magenta horizontal bars represent the median.



**Extended Data Figure 3 | Trans-synaptic labelling of dimorphic connections using iBLINC.** **a**, Overview of synaptic connections labelled in this paper (Fig. 1 and this figure). Some connections were labelled with both GRASP and iBLINC, yielding similar results. We generally note that the number of synapses is roughly reproducible from animal to animal and the number of the fluorescent dots is roughly comparable to the number of synapses identified by the electron microscopy analysis. However, there is also some variance from animal to animal (quantified in Fig. 3), consistent with previous analysis<sup>34</sup>. **b**, Labelling data not shown in Fig. 1. Fluorescent micrographs of the preanal ganglion region of transgenic

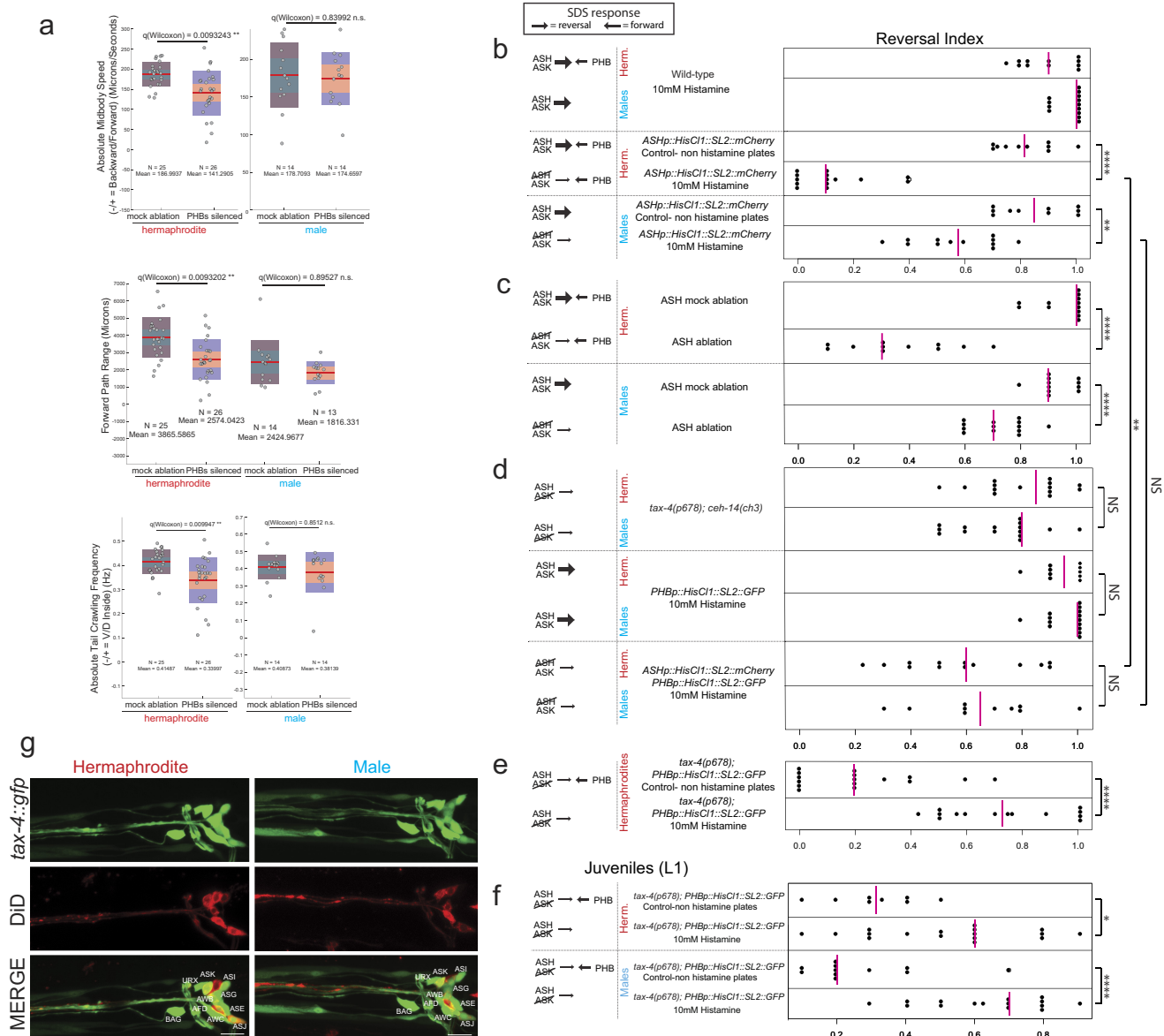
animals expressing the presynaptic *BirA::nrx-1* fusion in PHB (using the *gpa-6* promoter), AVG (using the *inx-18* promoter) and PHA (using the *srg-13* promoter), and postsynaptic acceptor peptide::*nlg-1* fusion in AVG and DA9 (using the *acr-2* promoter). Transgenic worms also express the streptavidin detector fused to 2 × sfGFP from the coelomocytes (*unc-122* promoter)<sup>10</sup>. Neuronal processes are labelled with cytoplasmic Cherry markers of the iBLINC pairs. Region of neurite overlap and observed synaptic puncta are marked with white boxes. Scale bars, 10 µM. Anterior is left and dorsal is up.



**Extended Data Figure 4 | Specificity of driver lines.** **a**, A 2.6 kb *gpa-6* promoter fragment fused to GFP is expressed consistently in PHB at all stages. There is also faint and variable expression in AWA. **b**, The 3.1 kb *flp-18* promoter fused to GFP is expressed consistently in AVA, and dimly and variably in AIY (85% of animals) and RIM (15% of animals), which were identified based on comparison to published *flp-18* expression

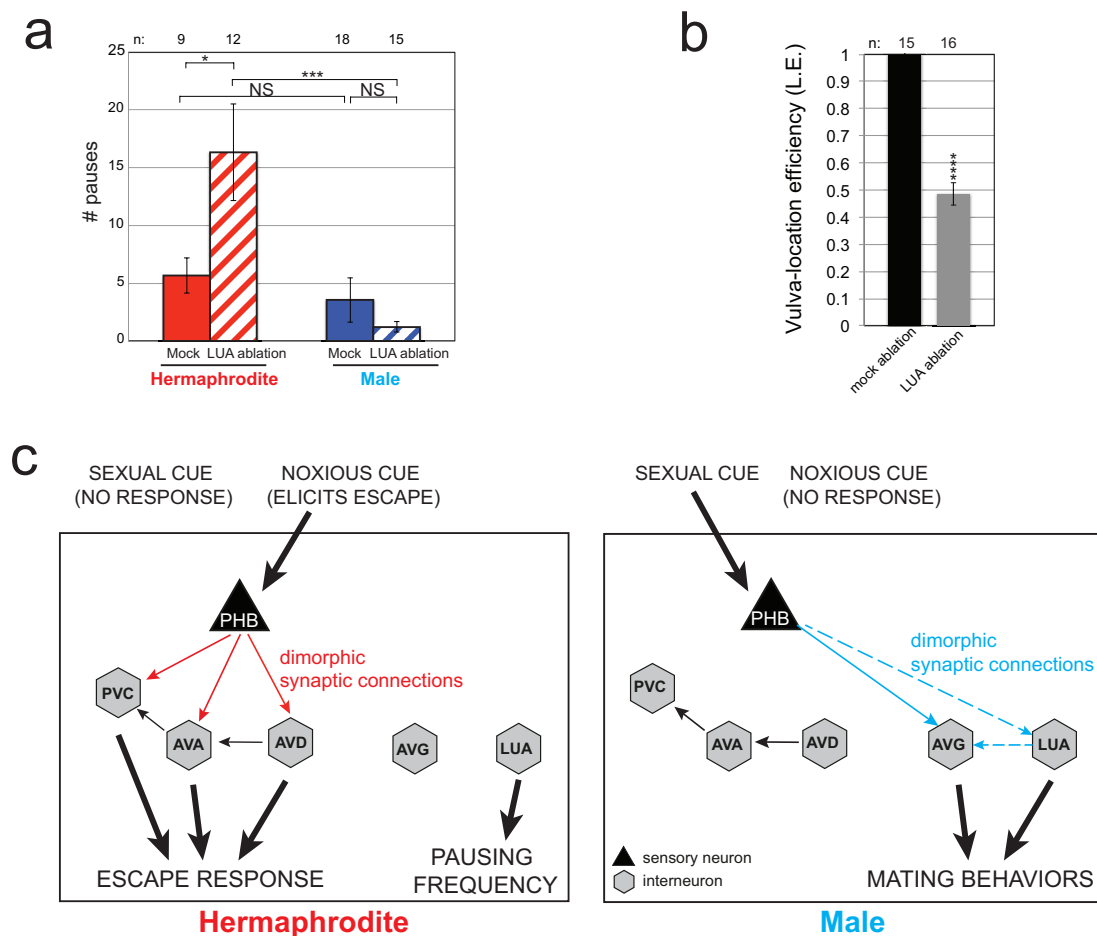
patterns<sup>35</sup>. **c**, The 1.8 kb *inx-18* 2nd intron fused to codon-optimized Cherry is expressed brightly and consistently in AVG, and dimly and variably in URXs. The AIY motif present in this fragment was deleted (see Methods). **d**, The 170 bp *eat-4* promoter (*eat-4p9*) fused to GFP is expressed in LUAs and PVR.





**Extended Data Figure 5 | Additional SDS avoidance assays.** **a**, PHB silenced hermaphrodites move slower forward (quantified as absolute mid-body speed) and as a result cover less of the plate (quantified as forward path range) compared to control hermaphrodites, whereas PHB-silenced males do not show any difference compared to control males. In addition, the tail-bending wave is affected in PHB-silenced hermaphrodites, but not in males (quantified as tail crawling frequency). Statistics were computed using Wilcoxon rank-sum test, and correction for multiple testing ( $q$  values) was computed across all measures (approximately 1404 tests) using the Benjamini–Hochberg procedure. **b**, Silencing of ASH neuronal activity using the histamine chloride channel 1 (*HisCl1*) affects the animals' chemosensory avoidance response. Males and hermaphrodites were assayed for effects of histamine on SDS avoidance behaviour. We used the *him-5* mutant background (which gives a high incidence of male progeny) as wild type. There is no difference between worms assayed in the presence and absence of histamine (Fig. 2a). The avoidance index of single animals was calculated as the fraction of reversal responses in 10 or more assays, depicted as black dots. Magenta vertical bars represent the median. L4 animals carrying the *kyEx5104* (*pNP424* (*sra-6::HisCl1::SL2::mCherry*))<sup>11</sup> transgene were grown on 10 mM histamine-containing NGM plates for 24 h. As a control, *kyEx5104* animals were grown on NGM plates without histamine. ASH silencing reduces the head sensory response to SDS, thus in hermaphrodites the antagonizing activity of the PHBs inhibits the backward movement and the worms do not reverse. In males, no such antagonizing activity occurs and the worms reverse, albeit with reduced ability. **c**, Ablation of ASH neurons affects the animals' chemosensory avoidance response in a similar manner to histamine-induced silencing. *sra-6:gfp* was used to

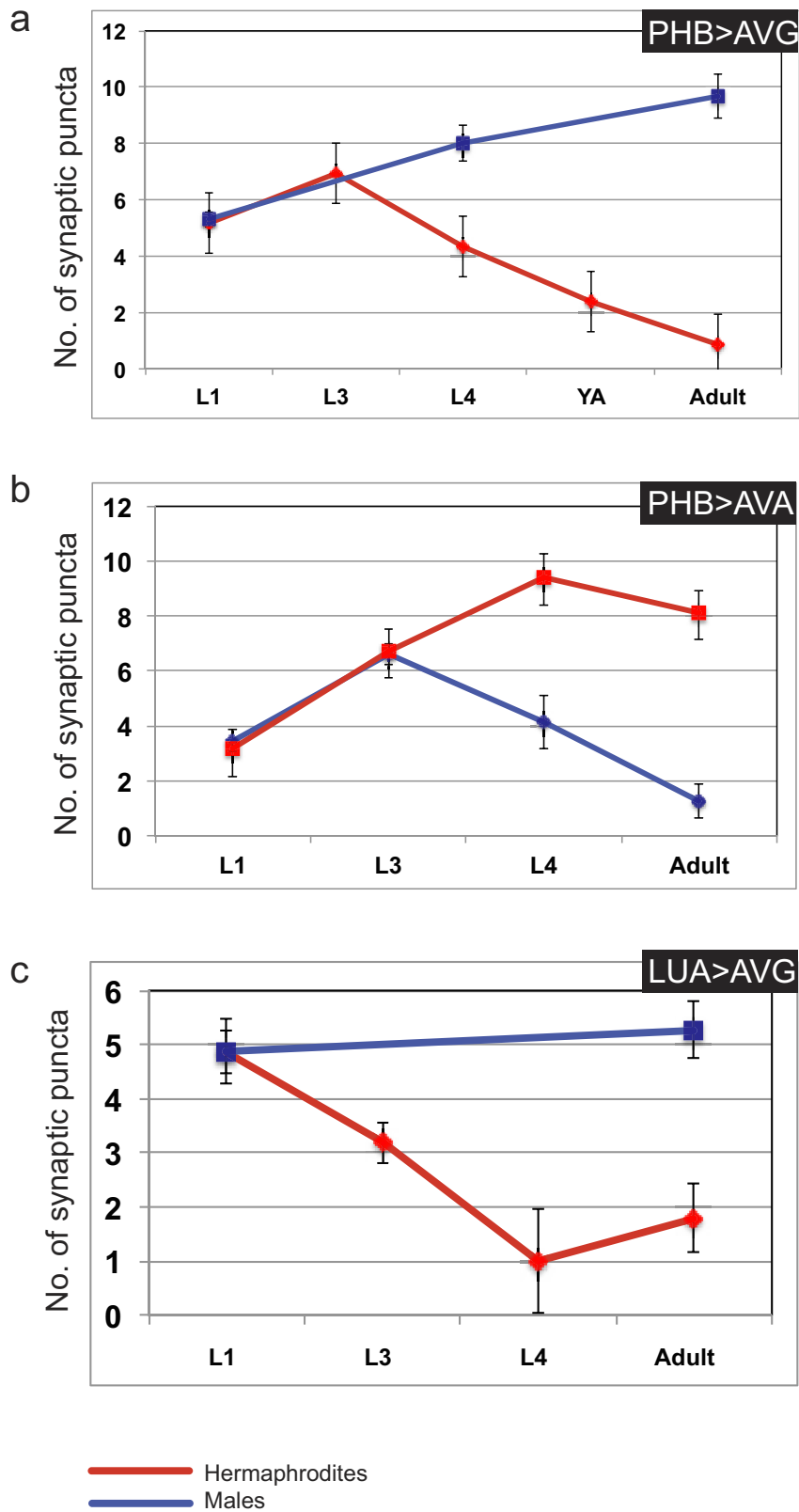
identify the ASH neurons. **d**, Behavioural differences stem from dimorphic connectivity differences and not from amphid/phasmid sensory function. *tax-4*; *ceh-14* double mutants behave in a similar manner in both sexes. PHB silencing (*gpa-6p::HisCl1::SL2::gfp*) does not affect behaviour in either sex. Silencing both the ASHs and PHBs in males showed no difference compared to ASH-silenced males. However, silencing ASHs and PHBs in hermaphrodites showed a significant difference compared to ASH silenced hermaphrodites, where we expect the PHBs to function in an antagonistic manner; thus in its absence, ASH-silenced hermaphrodites showed an increased ability to respond to SDS by reversing. **e**, PHB silencing in *tax-4* mutant background. *tax-4* is a subunit of a cyclic nucleotide gated channel expressed in chemosensory and thermosensory neurons<sup>36</sup>, see **g**. *tax-4* animals show a strongly reduced avoidance response to SDS<sup>12</sup>. Silencing of PHBs in *tax-4* hermaphrodites eliminated the antagonizing affect and animals were able to avoid SDS by backing. **f**, PHB silencing in *tax-4* mutant background at the L1 stage. Lack of avoidance seen in *tax-4* L1 males and hermaphrodites depends on PHB function. For all panels, we performed the nonparametric Mann–Whitney test with Bonferroni correction for multiple comparisons. \*\*\*\* $P < 0.0001$ , \*\* $P < 0.01$ , \* $P < 0.05$ ; NS, not significant. **g**, *tax-4* expression pattern is identical in hermaphrodites and males. *kyEx744* (*tax-4p::TAX-4::gfp*)<sup>36</sup>, was analysed in adult male and hermaphrodites. Amphid neurons in the head (ADL, ASH, ASI, ASJ, ASK, AWB) were stained using DiD to facilitate cell identification. Neurons identified, shown in the 'Merge' panel are identical in both sexes and match published data. All neurons are bilaterally symmetric left–right pairs, and for simplicity only left cells are shown. Scale bars, 10 μm.



#### Extended Data Figure 6 | Additional behavioural analysis.

**a**, Hermaphrodites in which LUAs have been ablated pause more frequently than mock-ablated hermaphrodites and LUA-ablated males. Error bars, s.e.m. **b**, LUA laser-ablated animals tested for the male's vulva location efficiency. The behavioural data shown in **a** and **b** supports the reported connectivity data shown in Extended Data Fig. 2. We performed

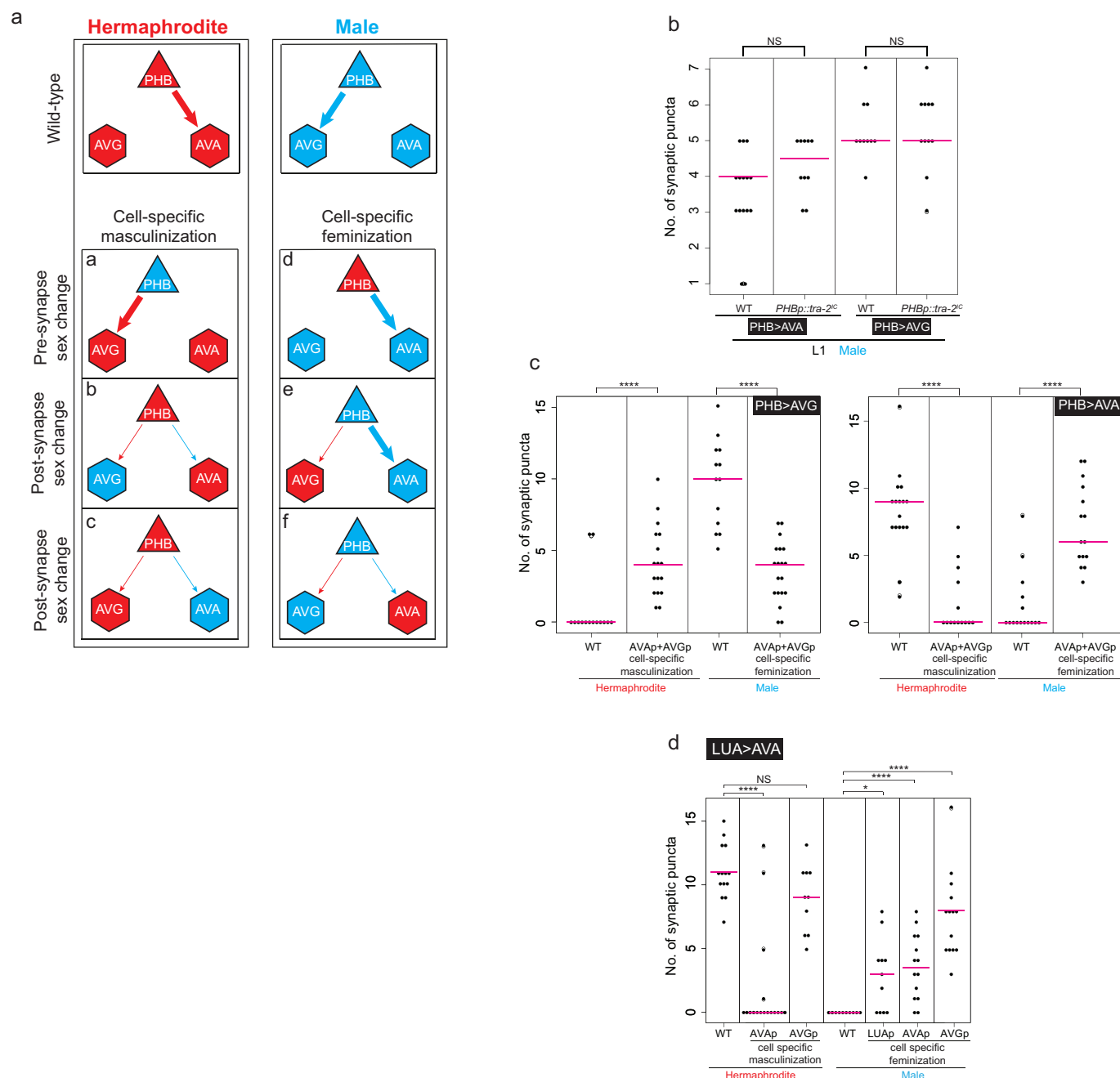
the nonparametric Mann–Whitney test (Wilcoxon rank sum test) with Bonferroni correction for multiple comparisons (**a**, **b**). Error bars, s.e.m. (**b**). \*\*\*\* $P < 0.0001$ , \*\*\* $P < 0.001$ , \* $P < 0.05$ ; NS, not significant. **c**, Summary of dimorphic behaviours induced by PHB sensory neurons and AVG/LUA/AVA interneurons.



**Extended Data Figure 7 | Time-course analysis of synapse pruning and development.** Hermaphrodites and males were analysed at the L1, L3, L4, young adult and gravid adult stages, and the number of synaptic puncta observed at each stage was plotted against developmental time points. Synaptic puncta in hermaphrodites are plotted in red, synaptic

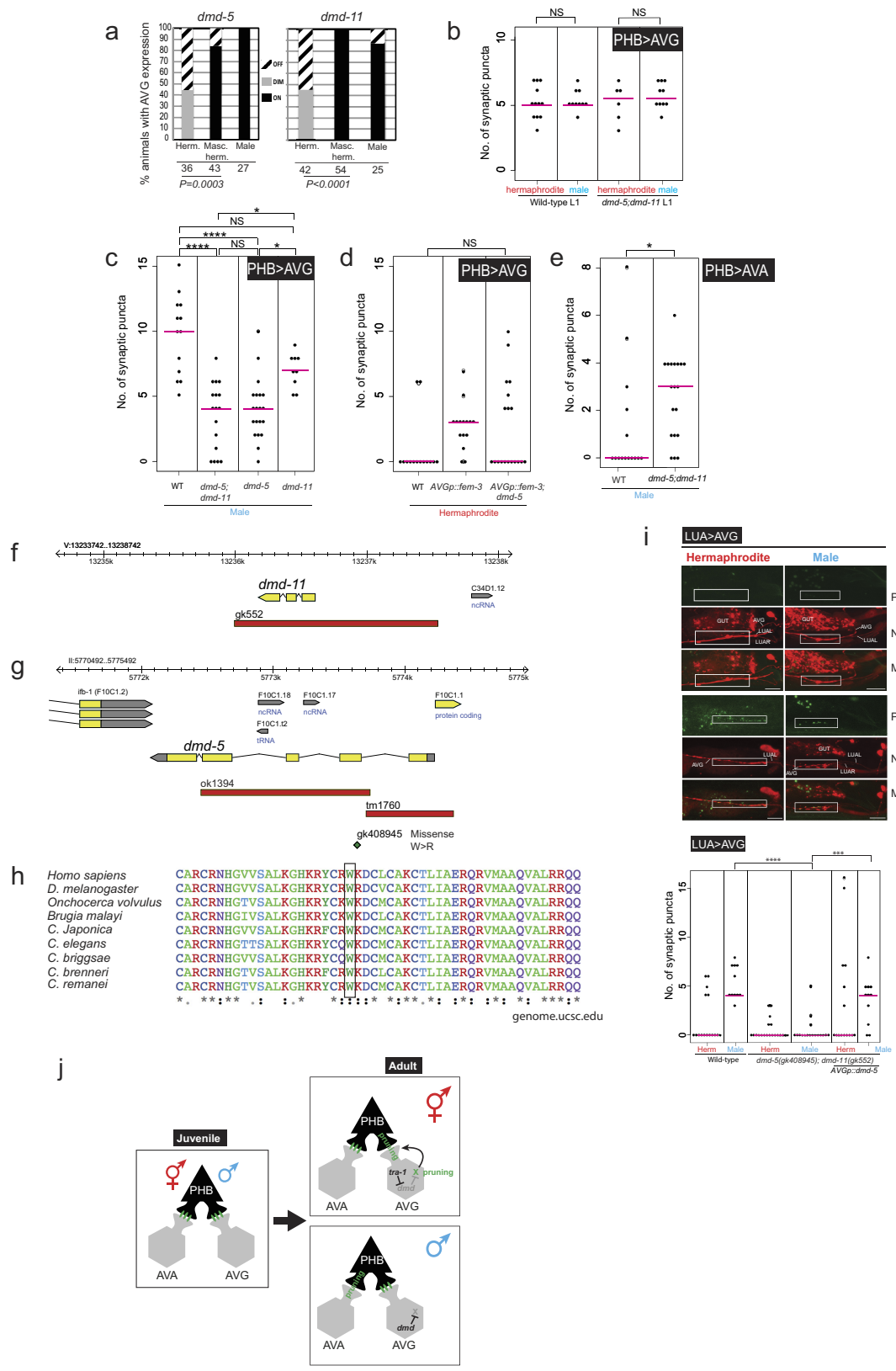
puncta in males are plotted in blue. **a**, PHB–AVG synapses are pruned in hermaphrodites at the L3 stage. **b**, PHB–AVA synapses are pruned in males at the L3 stage. **c**, LUA–AVG synapses are pruned earlier, starting at the L1 stage in hermaphrodites. Error bars, s.e.m.; for each time point depicted in graphs, at least 15 animals were analysed.





**Extended Data Figure 8 | Autonomy and non-autonomy of sex-specific synapse pruning.** **a**, Cartoon summarizing sex-change effects on synapses. **b**, L1-stage connectivity is not affected by sex reversal. **c**, Simultaneous sex reversal of both AVA and AVG. **d**, Masculinization of the postsynaptic cell AVA is sufficient to induce LUA–AVA synaptic puncta in hermaphrodites. Masculinization of AVG was not sufficient to induce synapses between

LUA and AVA in hermaphrodites. Feminizing LUA, AVA and AVG by expression of TRA-2<sup>LC</sup> was sufficient to induce ectopic LUA–AVA puncta in males. We performed the nonparametric Mann–Whitney test with Bonferroni correction for multiple comparisons. \*\*\*\* $P < 0.0001$ , \* $P < 0.01$ ; NS, not significant. Magenta horizontal bars represent the median.



Extended Data Figure 9 | See next page for caption.

**Extended Data Figure 9 | *dmd-5* and *dmd-11* expression, sequence and function.** **a**, Quantification of dimorphic expression of *dmd-5* and *dmd-11* in AVG. Expression in hermaphrodites was off or extremely faint. Expression of *inx-18p::FEM-3* derepressed *dmd-5* and *dmd-11* gene expression in hermaphrodite AVGs. Statistics calculated using Fischer's exact test. **b**, Quantification of the number of PHB–AVG synaptic puncta in L1 *dmd-5(gk408945); dmd-11(gk552)* double mutants, compared with wild-type L1 animals. At the L1 stage, *dmd-5* and *dmd-11* do not affect PHB–AVG synapses, suggesting they are required for maintenance of mature synapses. **c**, *dmd-5* single mutants and *dmd-5; dmd-11* double mutants display similar alterations in AVG synaptic wiring. **d**, *dmd-5* mutation suppresses the ectopic PHB–AVG synapses in AVG-masculinized animals. **e**, The PHB–AVA connection is non-autonomously partially stabilized in *dmd-5; dmd-11* mutants. **f**, *dmd-11* genomic locus and *gk552* deletion location. **g**, *dmd-5* genomic locus and mutation description. *ok1394* location was not curated, to determine location we used the following primers: forward primer: 5'-CAGAATGCCTGTTTCTCCGTC-3'; and reverse: 5'-CACTGCTTTTCCCGTTCAAAC-3'. *ok1394* and *tm1760* were both found to have an embryonic lethal phenotype that could not be rescued by the genomic locus (data not shown); thus we searched for single-point mutations of the 'million mutation project'<sup>37</sup>. *gk408945* is a missense substitution mutation of W54 to R, located in the second exon. Genomic analysis revealed that this mutation lies within the conserved DM domain (**h**), with perfect conservation across evolution. **h**, DM-domain

sequence conservation and location of *gk408945* mutation. Conservation and multiple sequence alignment were performed using UCSC Genome Browser (<http://genome.ucsc.edu>) and ClustalW. The DM domain is an intertwined zinc-containing DNA binding module. The DM domain binds DNA as a dimer, allowing the recognition of pseudopalindromic sequences<sup>38</sup>. **i**, *dmd-5* and *dmd-11* are required for maintenance of AVG synapses. Fluorescent micrographs and quantification of synaptic puncta of LUA–AVG. Region of neurite overlap and observed synaptic puncta marked with white boxes. M, merge; N, neurite; P, puncta. Statistics were calculated using the nonparametric Mann–Whitney test (**b**, **d**, **e**, **i**) or Kruskal–Wallis test with Dunn's multiple comparison test (**c**). \*\*\*\* $P < 0.0001$ , \*\*\* $P < 0.001$ , \* $P < 0.05$ ; NS; not significant. Magenta horizontal bars represent the median. When using a parametric *t*-test, there is also a significant difference for the LUA–AVG synapse between *dmd-5;dmd-11* mutant hermaphrodites and *dmd-5;dmd-11* mutant hermaphrodites that overexpress DMD-5 (\* $P < 0.05$ ). **j**, Summary of data. TRA-1 and DMD proteins are commonly thought to work as transcriptional repressors<sup>22</sup>. As *dmd-5* and *dmd-11* are already dimorphically expressed in AVG in embryos and L1 stage animals (not shown in this schematic), there must be other timer mechanisms that control the onset of pruning. For example, DMD-5 and DMD-11 may work together with a regulatory factor of the stage-specifically acting heterochronic pathway. Furthermore, we hypothesize that other neurons, such as the AVA neuron, may have its own complement of sex-specific *dmd* genes that control pruning.



# Interconnected microbiomes and resistomes in low-income human habitats

Erica C. Pehrsson<sup>1\*</sup>, Pablo Tsukayama<sup>1\*</sup>, Sanket Patel<sup>1,2</sup>, Melissa Mejía-Bautista<sup>1,3</sup>, Giordano Sosa-Soto<sup>1,3</sup>, Karla M. Navarrete<sup>3</sup>, Maritza Calderon<sup>4</sup>, Lilia Cabrera<sup>5</sup>, William Hoyos-Arango<sup>3</sup>, M. Teresita Bertoli<sup>3</sup>, Douglas E. Berg<sup>6,7</sup>, Robert H. Gilman<sup>4,5,8</sup> & Gautam Dantas<sup>1,2,6,9</sup>

**Antibiotic-resistant infections annually claim hundreds of thousands of lives worldwide. This problem is exacerbated by exchange of resistance genes between pathogens and benign microbes from diverse habitats. Mapping resistance gene dissemination between humans and their environment is a public health priority. Here we characterized the bacterial community structure and resistance exchange networks of hundreds of interconnected human faecal and environmental samples from two low-income Latin American communities. We found that resistomes across habitats are generally structured by bacterial phylogeny along ecological gradients, but identified key resistance genes that cross habitat boundaries and determined their association with mobile genetic elements. We also assessed the effectiveness of widely used excreta management strategies in reducing faecal bacteria and resistance genes in these settings representative of low- and middle-income countries. Our results lay the foundation for quantitative risk assessment and surveillance of resistance gene dissemination across interconnected habitats in settings representing over two-thirds of the world's population.**

Antibiotic resistance in bacterial pathogens causes hundreds of thousands of annual fatalities globally<sup>1–3</sup>. The spread of resistant organisms and their antibiotic resistance genes occurs by direct contact between humans and via interactions with environmental microbiota<sup>4–6</sup>. Horizontal gene transfer (HGT) facilitates this dissemination, and the reservoir of antibiotic resistance genes (the ‘resistome’<sup>7</sup>) in the environment, from which pathogens could theoretically draw, is ancient, diverse, and widespread<sup>8–10</sup>. Characterizing resistome distributions and their potential for dissemination across diverse habitats can identify the microbiota and antibiotic resistance genes that pose the highest risks to human health.

Most resistome studies have focused on either industrialized<sup>11–13</sup> or remote, ‘pristine’ settings<sup>8,9,14</sup>. However, most of the world's people reside outside these extremes: ~5.8 billion live in low- and middle-income countries<sup>15</sup>, with 863 million people living in slums<sup>16</sup>. Here, we characterized the microbiomes and resistomes of human faecal and co-localized, ecologically diverse environmental microbiota from two low-income, resource-limited Latin American settings: (1) a rural village of subsistence farmers in El Salvador (RES) and (2) a peri-urban shanty-town (slum) in Lima, Peru (PST), which represent critically understudied microbial ecosystems (Extended Data Fig. 1a). Aspects of life in such settings that are distinct from both hunter-gatherer and industrialized populations include crowding (in PST), limited access to clean drinking water and sanitation, supplementation of personally grown produce and livestock with processed foods, and ready access to antibiotics without prescriptions<sup>17</sup>. Furthermore, such industrializing countries are responsible for the majority of the worldwide 36% increase in antibiotic use between 2000 and 2010 (ref. 18), making investigation of antibiotic resistance transfer in these settings a global public health priority.

We analysed 263 faecal samples from 115 individuals in 27 houses over two years from RES and PST, as well as 209 environmental samples from donor households and surrounding areas in these communities.

The environmental samples included faeces from domesticated animals, soil, water, and samples from the sanitation facilities of each community: composting latrines in RES and a district-wide sewage system with a modern wastewater treatment plant (WWTP) in PST. We used a combination of 16S sequencing<sup>10,19</sup>, high-throughput functional metagenomic selections<sup>10,20,21</sup>, and whole-metagenome shotgun sequencing<sup>11,22</sup> to compare the phylogenetic architectures of these microbial populations and their associated resistomes.

## Resistome correlates with phylogeny across habitats

Small-insert metagenomic expression libraries in *Escherichia coli* constructed from 51 human faecal and 45 environmental samples from RES and PST (representing 258 Gb) were selected for functional resistance against 17 antibiotics (Supplementary Table 1). Sequencing and annotation<sup>4,23</sup> of these selections identified 1,100 unique (100% amino acid identical) encoded antibiotic resistance proteins collectively conferring resistance against all antibiotics except meropenem (see Methods; Supplementary Table 2). A total of 121 of these proteins were novel (<60% amino acid identity to any protein in NCBI nr), the majority of which (72%) were predicted antibiotic modifiers, including 57 class A  $\beta$ -lactamases. RES latrine libraries yielded the most novel proteins (46%), proportionally more than expected (Pearson's chi-squared test,  $P < 2 \times 10^{-5}$ ; Extended Data Fig. 1c).

To further characterize RES and PST resistome diversity and abundance, we performed whole-metagenome shotgun sequencing on 191 human faecal and 94 environmental samples (representing 344 Gb; Extended Data Fig. 1b; see Methods). We used ShortBRED<sup>24</sup> to quantify translated antibiotic resistance gene abundance in all sequenced metagenomes using a custom antibiotic resistance database that included antibiotic resistance genes identified here (see Methods; Supplementary Tables 3–5). RES and PST human-associated and environmental resistomes were related along an ecological gradient

<sup>1</sup>Center for Genome Sciences and Systems Biology, Washington University School of Medicine, St Louis, Missouri 63110, USA. <sup>2</sup>Department of Pathology and Immunology, Washington University School of Medicine, St Louis, Missouri 63110, USA. <sup>3</sup>Facultad de Ciencias de la Salud “Dr. Luis Edmundo Vásquez”, Universidad Dr. José Matías Delgado, El Salvador. <sup>4</sup>Laboratorios de Investigación y Desarrollo, Universidad Peruana Cayetano Heredia, San Martín de Porres, Lima 31, Peru. <sup>5</sup>Asociación Benéfica PRISMA, San Miguel, Lima 32, Peru. <sup>6</sup>Department of Molecular Microbiology, Washington University School of Medicine, St Louis, Missouri 63110, USA. <sup>7</sup>Department of Medicine, University of California San Diego, La Jolla, California 92093, USA. <sup>8</sup>Department of International Health, Johns Hopkins School of Public Health, Baltimore, Maryland 21205, USA. <sup>9</sup>Department of Biomedical Engineering, Washington University, St Louis, Missouri 63105, USA. \*These authors contributed equally to this work.

in terms of input from human faeces (Bray–Curtis; Extended Data Fig. 2a), with habitat explaining 22.4% of resistome variation (adonis,  $P < 0.001$ , Bray–Curtis).

A similar ecological gradient by habitat was observed when considering phylogenetic composition, based on 16S sequencing of 228 human faecal and 203 environmental samples from RES and PST (Extended Data Fig. 1b, see Methods), with habitat explaining even more of the variation between samples (41.9%; adonis,  $P < 0.001$ , weighted UniFrac; Extended Data Fig. 2b). Procrustes analysis confirmed that antibiotic resistance is significantly correlated with community composition (Bray–Curtis,  $M^2 = 0.360$ ,  $P < 0.001$ ; Extended Data Fig. 2c), not randomly distributed across habitats<sup>10,23</sup>.

### RES/PST versus global human faecal microbiota

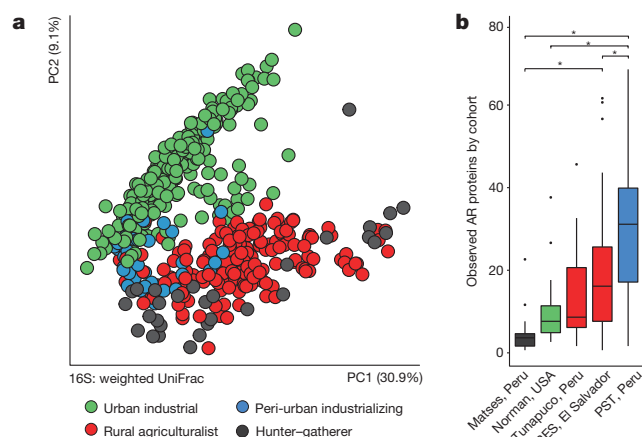
The human gut microbiota is highly diverse globally, with phylogenetic and functional variation potentially driven by age, diet, cultural traditions, pathogen carriage, and periodic perturbation (for example, by antibiotic exposure)<sup>19,22</sup>. We compared the RES and PST human faecal microbiota to published microbiota data sets from the USA, Malawi, Venezuela, and Peru (Amazonian hunter–gatherers and Andean highlands agriculturalists) (see Methods)<sup>14,19,25</sup>, classifying PST as ‘peri-urban industrializing’ and RES as ‘rural agriculturalist’. Faecal microbiota composition clustered by host lifestyle<sup>19,25</sup>, despite differences in geographic origin and study: RES microbiota clustered with other rural agriculturalists and hunter–gatherers, while peri-urban PST microbiota were intermediate between these and the urban industrialized cohorts (adonis  $R^2 = 23.8\%$ ,  $P < 0.001$ , weighted UniFrac; Fig. 1a and Extended Data Fig. 3).

To evaluate whether antibiotic resistance burden correlated with industrialization<sup>14,26,27</sup>, we compared resistomes from RES and PST to industrialized USA, traditional hunter–gatherer, and rural agriculturalist communities from<sup>25</sup>, two of which are also Peruvian (Extended Data Fig. 4a–f; see Methods). PST had the greatest number of antibiotic resistance proteins per person, despite not being the most industrialized cohort analysed (Fig. 1b and Extended Data Fig. 4g), consistent with Chinese and Hadza hunter–gatherer faecal microbiota harbouring at least as much or more antibiotic resistance as Western industrialized microbiota<sup>11,22</sup>.

### RES human and environmental microbiota

Frequent contact with environmental reservoirs during subsistence farming and inadequate excreta management<sup>28</sup> likely promote antibiotic resistance exchange in rural settings. The RES community had only one source of filtered water, and contents of composting latrines were used to fertilize household agricultural plots after attempted sterilization, potentially recycling antibiotic resistance. Accordingly, we compared the microbiomes and resistomes of RES human faecal samples and their surroundings, including latrines, animal faeces, soils, and drinking water sources (see Methods).

The RES human faecal microbiota separated from soil and water along PC1 in a principal coordinate analysis (PCoA) visualization of their phylogenetic composition, and habitats differed at the phylum level (Fig. 2a, Extended Data Fig. 5a and Supplementary Tables 6–10). Latrines were equidistant to human faecal microbiota and soil (weighted UniFrac, non-parametric Student's  $t$ -tests with Bonferroni correction) and were enriched for Halomonadaceae, a family of halophilic organisms potentially selected by the alkaline latrine environment (Extended Data Fig. 5b, Supplementary Discussion). Cow and dog faecal microbiota, which are ecologically similar to human faecal microbiota but with greater environmental exposure, were also intermediate along PC1. Soil and water were closest to each other ( $P < 0.05$ , non-parametric Student's  $t$ -tests with Bonferroni correction) and varied along PC3 (6.3% of variation). Human faecal microbiota were more homogenous in phylogenetic composition than latrine, soil, and water microbiota, potentially because environmental samples encountered more diverse and variable conditions.



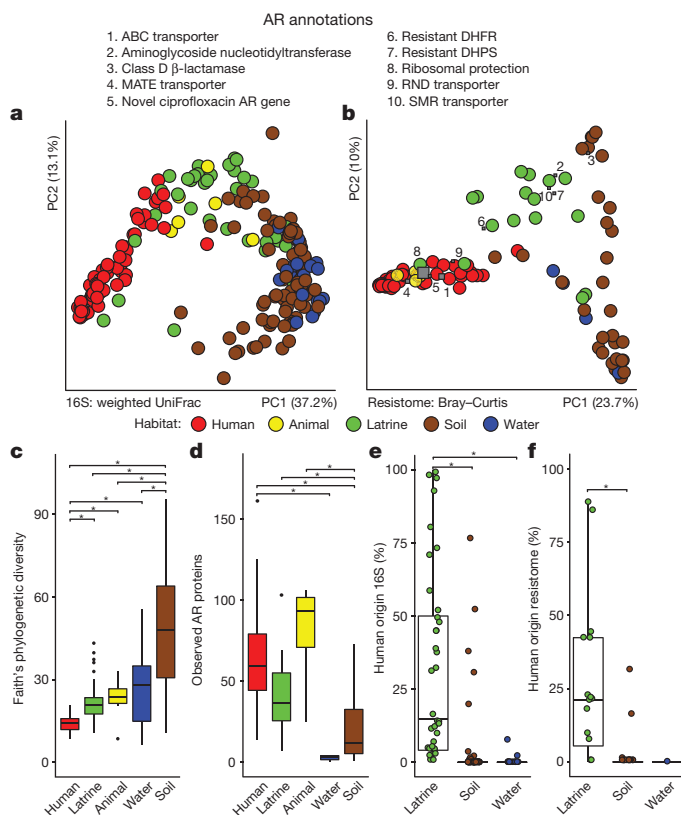
**Figure 1 | RES and PST human faecal microbiota and resistomes versus global populations.** **a**, PCoA of weighted UniFrac distances between RES ( $n = 60$ ) and PST ( $n = 46$ ) microbiota and published human faecal microbiota<sup>14,19,25</sup> ( $n = 446$ ; see Supplementary Table 14), coloured by host lifestyle. Adonis  $R^2 = 15.4\%$ ,  $P < 0.001$ . **b**, Number of antibiotic resistance proteins per person in RES ( $n = 42$ ) and PST ( $n = 44$ ) and published human faecal microbiota (ref. 25) ( $n = 53$ ; see Supplementary Table 15), coloured by host lifestyle. Error bars, s.d.; centre bars, median. \* $P < 0.05$ , non-parametric Student's  $t$ -tests, Bonferroni correction. AR, antibiotic resistance.

The RES resistomes exhibited a similar ecological gradient (Bray–Curtis, Fig. 2b). Although RES soil had the highest phylogenetic diversity (Fig. 2c), it contained fewer antibiotic resistance proteins per sample than all habitats but water (Fig. 2d). In contrast, human faecal microbiota had the lowest phylogenetic diversity, but more antibiotic resistance proteins per sample than both soil and water. However, soil and latrine resistomes were more heterogeneous than human and animal faecal resistomes. In particular, non-human RES resistomes were enriched in aminoglycoside nucleotidyltransferases, class D  $\beta$ -lactamases, SMR transporters, and resistant dihydropteroate synthetases and dihydrofolate reductases (Fig. 2b, Extended Data Fig. 5c, d; see Methods).

By analysing RES habitats prone to exchange with human faeces using SourceTracker<sup>29</sup>, we found that direct input from chicken faeces represents a potent avenue for the introduction of antibiotic resistance genes compatible with the human faecal microbiota into soil (see Methods). The contribution of RES human faecal resistomes to environmental resistomes (Fig. 2f) roughly recapitulated phylogenetic trends, with significantly higher input to latrines than soil or water (Fig. 2e; see Supplementary Information). Only the two soils collected adjacent to chicken coops had  $>1\%$  human faecal resistome input. These two samples clustered away from the other samples along PC3 (5.3% of variation) and were closer to human samples along PC1 in the PCoA visualization of all RES samples (Fig. 2a). Overall, RES human faecal microbiota were closer to soil collected near chicken coops than to any other soil location in both phylogenetic and antibiotic resistance composition (Bray–Curtis, non-parametric Student's  $t$ -tests with Bonferroni correction). They shared 80 antibiotic resistance proteins at  $>99\%$  identity, including three class C  $\beta$ -lactamases, which are common in Enterobacteriaceae, in contrast to two antibiotic resistance proteins shared between RES humans and soil from the pond edge. On average, human faecal and chicken coop soil resistomes shared 10 antibiotic resistance proteins, significantly more than with any other soil type (1–2 antibiotic resistance proteins, non-parametric Student's  $t$ -tests with Bonferroni correction).

### PST human and sewage microbiota

To monitor the impact of sewage treatment on resistomes, we sampled wastewater at street-level access points nearest to participating PST households and influent and effluent from the local WWTP and compared their phylogenetic composition and resistome to the

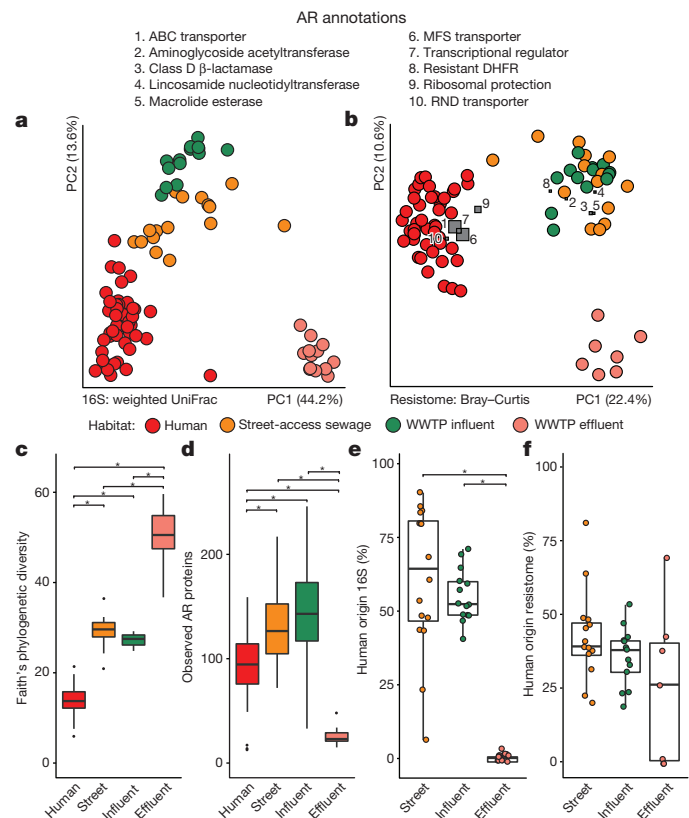


**Figure 2 | Salvadoran rural agriculturalist (RES) human faecal and environmental microbiota and resistomes.** **a, c, e,** Phylogenetic composition ( $n = 60, 6, 36, 84$  and  $22$  for human, animal, latrine, soil and water, respectively). **b, d, f,** Resistome ( $n = 42, 4, 16, 30$  and  $4$  for human, animal, latrine, soil and water, respectively). **a,** PCoA of weighted UniFrac distances between microbiota. Adonis  $R^2 = 43.2\%$ ,  $P < 0.001$ . **b,** PCoA of Bray–Curtis distances between resistomes with abundance-weighted coordinates of the top five most discriminative antibiotic resistance categories enriched in human or non-human habitats (squares, size proportional to overall abundance). Adonis  $R^2 = 26.6\%$ ,  $P < 0.001$ . **c, d,**  $*P < 0.05$ , non-parametric Student's  $t$ -tests, Bonferroni correction. **c,** Faith's phylogenetic diversity. **d,** Observed antibiotic resistance proteins. **e, f,** Percentage of latrine, soil, and water microbiota (**e**) and resistomes (**f**) attributable to human faeces, as determined by SourceTracker<sup>29</sup>.  $*P < 0.05$ , pairwise Wilcoxon tests, Bonferroni correction. Error bars, s.d.; centre bars, median.

faecal microbiota of PST residents (see Methods and Supplementary Information). Portions of treated wastewater effluent are discharged into the Pacific Ocean and also used to irrigate public parks and agricultural fields, potentially enabling re-introduction of antibiotic residues and antibiotic resistance genes into human communities<sup>30–32</sup>.

Although geographically closest to human donors, street-access sewage was more similar in microbial composition to WWTP influent (non-parametric Student's  $t$ -tests with Bonferroni correction,  $P < 0.001$ ), implying that even relatively brief periods in this non-human, aerobic environment caused a greater shift in bacterial composition than all changes downstream during transit through the sewage system. Overall, PST human waste underwent drastic changes in microbial composition as it progressed through sewage treatment, decreasing in similarity to human faecal microbiota at each subsequent stage (Student's non-parametric  $t$ -tests with Bonferroni correction,  $P < 0.001$ , weighted UniFrac; Fig. 3a, Extended Data Fig. 6a, b and Supplementary Table 11).

Sewage resistomes also decreased in similarity to PST human faecal resistomes at each treatment stage, although street-access sewage and WWTP influent were equally similar to human faeces in antibiotic resistance content (non-parametric Student's  $t$ -tests with Bonferroni

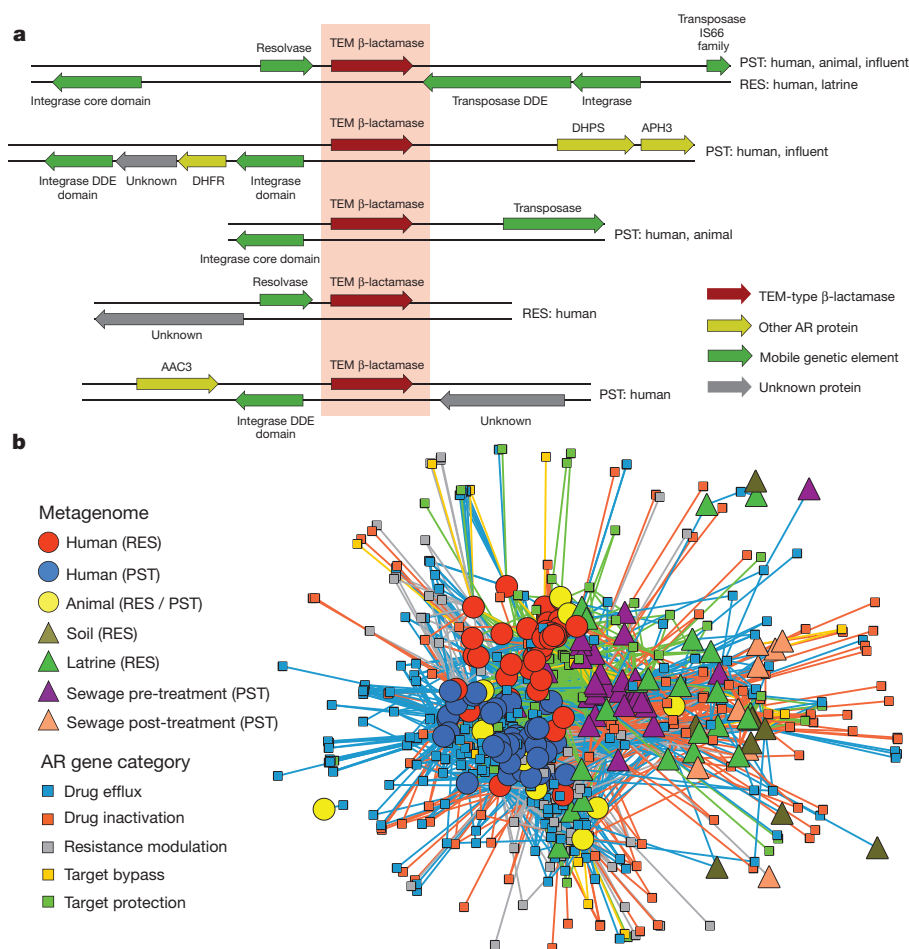


**Figure 3 | Peruvian peri-urban slum (PST) human faecal and sewage microbiota and resistomes.** **a, c, e,** Phylogenetic composition ( $n = 45, 16, 14$  and  $13$  for human, street-access, influent and effluent, respectively). **b, d, f,** Resistome ( $n = 44, 14, 13$  and  $7$  for human, street-access, influent and effluent, respectively). **a,** PCoA of weighted UniFrac distances between microbiota. Adonis  $R^2 = 58.0\%$ ,  $P < 0.001$ . **b,** PCoA of Bray–Curtis distances between resistomes with abundance-weighted coordinates of the top five most discriminative antibiotic resistance categories enriched in human or sewage habitats (squares, size proportional to overall abundance). Adonis  $R^2 = 32.3\%$ ,  $P < 0.001$ . **c, d,**  $*P < 0.05$ , non-parametric Student's  $t$ -tests, Bonferroni correction. **c,** Faith's phylogenetic diversity. **d,** Observed antibiotic resistance proteins. **e, f,** Percentage of sewage microbiota (**e**) and resistomes (**f**) attributable to human faeces at each sewage treatment stage, as determined by SourceTracker.  $*P < 0.05$ , pairwise Wilcoxon tests, Bonferroni correction. Error bars, s.d.; centre bars, median.

correction, Bray–Curtis; Fig. 3b). Although WWTP effluent had the highest phylogenetic diversity (Fig. 3c), it had the fewest antibiotic resistance proteins per sample (Fig. 3d), as observed for soil in RES. In contrast, street-access sewage and WWTP influent had both higher phylogenetic diversity and more antibiotic resistance proteins per sample than PST human faeces. Drug efflux antibiotic resistance mechanisms were higher overall in PST human faecal vs sewage resistomes ( $P < 0.05$ , pairwise Wilcoxon tests with Bonferroni correction; Fig. 3b and Extended Data Fig. 6c, d). In contrast, sewage was enriched for aminoglycoside acetyltransferases, class D  $\beta$ -lactamases, and resistant dihydrofolate reductases.

As expected, the human faecal contribution to WWTP effluent phylogenetic composition was lower than for street-access sewage or WWTP influent ( $P < 0.05$ , pairwise Wilcoxon tests with Bonferroni correction; Fig. 3e). Interestingly, this difference was not significant for resistomes, where human faecal input was high at all sewage treatment stages ( $P > 0.05$ ; Fig. 3f). Thus, although sewage treatment was successful in reducing the overall carriage of antibiotic resistance genes as well as depleting human faecal microbes, antibiotic resistance genes of faecal origin survived and could be reintroduced into the natural environments where effluent is discharged.





**Figure 4 | Antibiotic resistance proteins found in multiple habitats and genetic contexts in RES and PST. a**, Representative alignment of 5 of 25 contigs encoding a TEM-type  $\beta$ -lactamase at 99.9% nucleotide identity (full list of contigs in Methods). Contigs were annotated with Resfams v1.2. Source metagenomic libraries are indicated on the left. **b**, Antibiotic resistance networks between human and environmental metagenomes in RES and PST. Small nodes (squares) represent unique antibiotic resistance proteins found in at least one sampled metagenome, coloured by predicted resistance mechanism. Large nodes represent individual human/animal (circle) or environmental (triangle) metagenomes, coloured by habitat/cohort. Lines connecting samples and antibiotic resistance proteins represent a ShortBRED hit with an RPKM (reads per kilobase per million reads) of  $\geq 10$ , coloured by mechanism.

As the presence of antibiotics in sewage systems may influence resistome diversity and select for HGT, we used a modified solid phase extraction and mass spectrometry protocol<sup>33</sup> to detect 16 antibiotics from seven classes in 22 WWTP influent and effluent samples. Chloramphenicol, ciprofloxacin, tetracycline, trimethoprim, and sulfamethoxazole were consistently detected in influent throughout the sampling period, while erythromycin was detected in 36% of influent samples. Only sulfamethoxazole was detected in effluent samples (concentrations of 18–26,000  $\mu\text{g l}^{-1}$ ) (Supplementary Table 12). Antibiotics detected in WWTP samples were among the highest selling antibiotics in Peru<sup>18</sup> and may enrich for antibiotic resistance in these bacterial communities and in those exposed to effluent. However, no  $\beta$ -lactams were detected in any WWTP sample, despite the high abundance of  $\beta$ -lactam resistance genes found in our metagenomic surveys and amoxicillin being the highest-selling antibiotic in Peru<sup>18</sup>. This suggests that  $\beta$ -lactam antibiotics are degraded to undetectable levels in humans or the sanitation system before reaching the WWTP.

### Highly cosmopolitan AR and HGT across microbiota

Although resistome and phylogenetic composition appear tightly linked in most microbial communities<sup>10,23</sup>, some clinically relevant antibiotic resistance proteins (for example, TEM, CTX-M, KPC, AAC-6') have been extremely successful in global dissemination via clonal expansion and HGT between multiple pathogen hosts<sup>34</sup>. We identified highly cosmopolitan antibiotic resistance proteins by comparing the prevalence of genes encoding them across all RES and PST habitats (Extended Data Fig. 7a). Two proteins, including the sulfonamide-resistant dihydropteroate synthetase (DHPS) Sul2, were found in 50% of samples in six of seven habitats. On our functional metagenomic contigs, ten of the twelve DHPSs were  $>98\%$  amino acid identical to Sul1, Sul2, or Sul3 and were co-localized with integrases and numerous

other antibiotic resistance genes, suggesting multidrug-resistance integrons may facilitate their broad distribution in these settings<sup>35</sup>.

To further investigate antibiotic resistance exchange potential in RES and PST, we examined flanking genetic sequences in the contig assemblies from our functional selections for evidence of past HGT. A total of 120 (11%) of our unique antibiotic resistance proteins were encoded in more than one genetic context (contigs with  $<90\%$  local nucleotide identity; Extended Data Fig. 7b), and the number of contexts was positively correlated with the number of metagenomic libraries (Spearman's  $\rho = 0.59$ ,  $P < 2.2 \times 10^{-16}$ ) and habitats (Spearman's  $\rho = 0.47$ ,  $P < 2.2 \times 10^{-16}$ ) in which an antibiotic resistance protein was encoded. One TEM  $\beta$ -lactamase (TEM-1) was encoded in 25 contexts (Fig. 4a). In contrast, 41% of antibiotic resistance proteins found in multiple habitats were always encoded in the same genetic context. For instance, a TetX (a tetracycline-inactivating enzyme) variant was encoded in the same context in human and animal faeces, latrines, and sewage influent. Three of the six antibiotic resistance proteins encoded by both human faecal and soil microbiota were encoded in the same genetic context (CblA and TEM class A  $\beta$ -lactamases and a class D  $\beta$ -lactamase).

We revealed a large network of antibiotic resistance gene sharing between microbial communities of human, animal, and environmental origin (Fig. 4b), facilitated by HGT between bacterial genomes and spread of bacterial hosts across communities. To further assess the potential mobility of the antibiotic resistance genes found in our contigs, we identified putative mobile genetic elements (MGEs) and multidrug resistance clusters (MDRCs) by annotation (Supplementary Table 13; see Methods). There was a small but significant positive correlation between the proportion of antibiotic resistance contigs with an MGE or MDRC and the number of libraries and habitats in which the antibiotic resistance protein was encoded (Spearman's  $\rho = 0.11$ –0.17,

$P < 4.3 \times 10^{-4}$ ), which supports a role for MGEs and MDRCs in antibiotic resistance transfer across environments and increased accessibility to pathogens<sup>36</sup>. When ecological analyses were restricted to only antibiotic resistance genes adjacent to an MGE on one of our functional contigs, the results largely recapitulated the trends observed with the full antibiotic resistance set, with resistome correlating with phylogenetic composition across ecological habitats. Additionally, the RES human faecal contribution was reduced in soil compared to latrine resistomes, while the contribution of PST human faeces to sewage was not significantly different before or after treatment (Supplementary Discussion and Extended Data Fig. 8).

## Conclusions

Our characterization of human faecal and environmental microbiota and their resistomes from two low-income settings in Latin America is particularly relevant to global public health. Billions of people currently live in rural or transitional areas around large urban centres, where unregulated access to antibiotics and limited access to clean water and improved sanitation increase the risk of pathogen transmission. Future studies on the factors that promote or restrict antibiotic resistance exchange between environmental microbiota, human commensals, and pathogens, particularly during waste treatment, are merited. These would involve real-time molecular surveillance of 'high-risk' environments (for example, hospitals, large-scale animal feeding operations) to identify specific routes for the spread of resistant bacteria and antibiotic resistance genes and inform the design of public health interventions to decrease their global enrichment and dissemination.

**Online Content** Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 5 January; accepted 16 March 2016.

1. *Antimicrobial Resistance: Global Report on Surveillance* 1st edn (World Health Organization, 2014).
2. Centers for Disease Control and Prevention. *Antibiotic Resistance Threats in the United States, 2013* (2013).
3. The Review on Antimicrobial Resistance. *Antimicrobial Resistance: Tackling a Crisis for the Health and Wealth of Nations* (2014).
4. Forsberg, K. J. *et al.* The shared antibiotic resistome of soil bacteria and human pathogens. *Science* **337**, 1107–1111 (2012).
5. Allen, H. K. *et al.* Call of the wild: antibiotic resistance genes in natural environments. *Nature Rev. Microbiol.* **8**, 251–259 (2010).
6. Martinez, J. L. The role of natural environments in the evolution of resistance traits in pathogenic bacteria. *Proc. R. Soc. Lond. B* **276**, 2521–2530 (2009).
7. Wright, G. D. The antibiotic resistome: the nexus of chemical and genetic diversity. *Nature Rev. Microbiol.* **5**, 175–186 (2007).
8. D'Costa, V. M. *et al.* Antibiotic resistance is ancient. *Nature* **477**, 457–461 (2011).
9. Allen, H. K., Moe, L. A., Roddumr, J., Gaarder, A. & Handelsman, J. Functional metagenomics reveals diverse  $\beta$ -lactamases in a remote Alaskan soil. *ISME J.* **3**, 243–251 (2009).
10. Forsberg, K. J. *et al.* Bacterial phylogeny structures soil resistomes across habitats. *Nature* **509**, 612–616 (2014).
11. Hu, Y. *et al.* Metagenome-wide analysis of antibiotic resistance genes in a large cohort of human gut microbiota. *Nature Commun.* **4**, 2151 (2013).
12. Li, B. *et al.* Metagenomic and network analysis reveal wide distribution and co-occurrence of environmental antibiotic resistance genes. *ISME J.* **9**, 2490–2502 (2015).
13. Smillie, C. S. *et al.* Ecology drives a global network of gene exchange connecting the human microbiome. *Nature* **480**, 241–244 (2011).
14. Clemente, J. C. *et al.* The microbiome of uncontacted Amerindians. *Science Advances* **1**, e1500183 (2015).
15. The World Bank Group. *Data: Countries: High Income* (<http://data.worldbank.org/income-level/HIC>) (2015).
16. World Health Organization. *Global Health Observatory (GHO) Data: Urban Health* ([http://www.who.int/gho/urban\\_health/en/](http://www.who.int/gho/urban_health/en/)) (2015).
17. Okeke, I. N. *et al.* Antimicrobial resistance in developing countries. Part I: recent trends and current status. *Lancet Infect. Dis.* **5**, 481–493 (2005).
18. Van Boeckel, T. P. *et al.* Global antibiotic consumption 2000 to 2010: an analysis of national pharmaceutical sales data. *Lancet Infect. Dis.* **14**, 742–750 (2014).
19. Yatsunenko, T. *et al.* Human gut microbiome viewed across age and geography. *Nature* **486**, 222–227 (2012).
20. Handelsman, J., Rondon, M. R., Brady, S. F., Clardy, J. & Goodman, R. M. Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. *Chem. Biol.* **5**, R245–R249 (1998).
21. Sommer, M. O., Dantas, G. & Church, G. M. Functional characterization of the antibiotic resistance reservoir in the human microflora. *Science* **325**, 1128–1131 (2009).
22. Rampelli, S. *et al.* Metagenome sequencing of the Hadza hunter-gatherer gut microbiota. *Curr. Biol.* **25**, 1682–1693 (2015).
23. Gibson, M. K., Forsberg, K. J. & Dantas, G. Improved annotation of antibiotic resistance determinants reveals microbial resistomes cluster by ecology. *ISME J.* **9**, 207–216 (2014).
24. Kaminski, J. *et al.* Fast and accurate metagenomic search with ShortBRED. *PLoS Comp. Biol.* **11**, e1004557 (2015).
25. Obregon-Tito, A. J. *et al.* Subsistence strategies in traditional societies distinguish gut microbiomes. *Nature Commun.* **6**, 6505 (2015).
26. Walson, J. L., Marshall, B., Pokhrel, B. M., Kafle, K. K. & Levy, S. B. Carriage of antibiotic-resistant fecal bacteria in Nepal reflects proximity to Kathmandu. *J. Infect. Dis.* **184**, 1163–1169 (2001).
27. Pallecchi, L. *et al.* Quinolone resistance in absence of selective pressure: the experience of a very remote community in the Amazon forest. *PLoS Negl. Trop. Dis.* **6**, e1790 (2012).
28. *Millennium Development Goals Report 2015* (United Nations, 2015).
29. Knights, D. *et al.* Bayesian community-wide culture-independent microbial source tracking. *Nature Methods* **8**, 761–763 (2011).
30. Stalder, T. *et al.* Quantitative and qualitative impact of hospital effluent on dissemination of the integron pool. *ISME J.* **8**, 768–777 (2013).
31. Baquero, F., Martinez, J. L. & Canton, R. Antibiotics and antibiotic resistance in water environments. *Curr. Opin. Biotechnol.* **19**, 260–265 (2008).
32. Munck, C. *et al.* Limited dissemination of the wastewater treatment plant core resistome. *Nature Commun.* **6**, 8452 (2015).
33. Li, B., Zhang, T., Xu, Z. & Fang, H. H. Rapid analysis of 21 antibiotics of multiple classes in municipal wastewater using ultra performance liquid chromatography-tandem mass spectrometry. *Anal. Chim. Acta* **645**, 64–72 (2009).
34. Hawkey, P. M. & Jones, A. M. The changing epidemiology of resistance. *J. Antimicrob. Chemother.* **64**, i3–i10 (2009).
35. Huovinen, P., Sundstrom, L., Swedberg, G. & Skold, O. Trimethoprim and sulfonamide resistance. *Antimicrob. Agents Chemother.* **39**, 279–289 (1995).
36. Martinez, J. L., Coque, T. M. & Baquero, F. What is a resistance gene? Ranking risk in resistomes. *Nature Rev. Microbiol.* **13**, 116–123 (2015).

**Supplementary Information** is available in the online version of the paper.

**Acknowledgements** We thank the residents of our study communities in El Salvador and Peru for their generosity and trust, without which this study would not have been possible; Epilogos Charities Inc. for on-site logistical support and community networking; the Fundación Luis Edmundo Vásquez (FUNDALV), Universidad Dr. José Matías Delgado, Asociación Benéfica Prisma, and Universidad Peruana Cayetano Heredia for logistical support in the collection and shipment of samples; S. del Pilar Basilio at SEDAPAL in Lima for facilitating access and sample collection at the 'PTAR San Juan' WWTP; J. Hoisington-Lopez at the Center for Genome Sciences and Systems Biology and staff at the Genome Technology Access Center at Washington University School of Medicine for generating Illumina sequencing data; S. Alvarez and staff at the Proteomics & Mass Spectrometry Facility at the Donald Danforth Plant Science Center for mass-spectrometry analyses of water samples; and members of the Dantas laboratory for discussions of the results and analyses. This work is supported in part by awards to G.D. through the Edward Mallinckrodt, Jr. Foundation (Scholar Award), the Children's Discovery Institute (MD-II-2011-117), and the National Institute of General Medical Sciences of the National Institutes of Health (R01-GM099538). Work at the DDPSC was supported by the National Science Foundation (DBI-0521250) for acquisition of the QTRAP LC-MS/MS instrument. E.C.P. is funded by the Department of Defense (DoD) through the National Defense Science and Engineering Graduate (NDSEG) Fellowship. The content is solely the responsibility of the authors and does not necessarily represent the official views of the funding agencies.

**Author Contributions** D.E.B., G.D., M.T.B., and E.C.P. planned the RES study; D.E.B., G.D., R.H.G., and P.T. planned the PST study; M.T.B. and W.H.A. implemented the RES study approval in El Salvador; E.C.P. implemented the RES study approval in the USA; R.H.G. and L.C. implemented the PST study approval in Peru; P.T. implemented the PST study approval in the USA; M.T.B., W.H.A., K.M.N., M.M.B., G.S.S., and E.C.P. collected surveys and samples in RES; P.T., M.C., and L.C. collected samples in PST; E.C.P., M.M.B., G.S.S., and S.P. extracted DNA and generated 16S, functional metagenomic, and shotgun data for RES samples; P.T. and S.P. extracted DNA and generated 16S, functional metagenomic, and shotgun data for PST samples; E.C.P. and P.T. performed analyses and interpreted results; and E.C.P., P.T., and G.D. wrote the paper with input from other co-authors.

**Author Information** Assembled functional metagenomic contigs and 16S and shotgun metagenomic reads have been deposited to NCBI GenBank and SRA (PRJNA300541). Reprints and permissions information is available at [www.nature.com/reprints](http://www.nature.com/reprints). The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to G.D. ([dantas@wustl.edu](mailto:dantas@wustl.edu)).

## METHODS

**Site and population overview.** The community sampled in Lima (PST) is a 'pueblo joven' (or shantytown) built on the slopes of the desert hills ~15 miles southwest of Lima, Peru (Extended Data Fig. 1a). The area was largely settled in the 1980s; the 2007 census recorded 56,915 inhabitants in an area of ~5 km<sup>2</sup>, a population density approximately four times higher than the rest of the Lima Metropolitan Area. Although accurate estimates of socioeconomic status were not available, the community is considered low income (although not uniformly) with an average family income of ~\$130 per month. Although the majority of the households have access to electricity, water and sewage, many households in the poorer hilltop settlements still lack access to these services. Most households are now linked to a district-wide sewage system that funnels waste from PST and neighbouring communities to the nearby 'San Juan' WWTP, one of 19 treatment plants serving the city of Lima. Sewage influent is collected in large aeration lagoons and subjected to stages of oxidation, settling, and chlorination in subsequent lagoons. Diarrheal diseases are common in this community, in particular among children, with an average of three episodes of diarrhoea per year due to high prevalence of various bacterial, viral and protozoan infectious agents<sup>37–39</sup>. Commonly used antibiotics are commonly purchased without prescription at the local pharmacies.

The rural Salvadoran village (RES) is home to approximately 100 people and is located in the mountains outside of San Salvador. Most individuals are subsistence farmers, although some are employed outside the village. They also participate in communal small-scale commercial enterprises, such as fish cultivation. Although the community has benefitted from infrastructure improvements, including pre-fabricated houses, drinking water is available only from a sand-filtered communal tap. Drinking water is stored in containers, and washing of dishes and clothing is primarily performed with unfiltered water or rainwater stored in outdoor barrels or reservoirs, which can be visibly contaminated with plant material. Each household has a double-vault composting latrine, a recommended method of waste disposal in low-income areas<sup>40</sup>, but which was available to only ~14% of rural Salvadorans in 2013 (ref. 41). Of the almost half (47%) of all people who live in rural areas worldwide<sup>42</sup>, 16% do not have drinking water sources protected from contamination with human excreta, and 50% lack sanitation facilities that separate excrement from human contact<sup>28</sup>. Urine is diverted away from the latrine, and wood ash is added to the latrine compartment after each use to increase the internal pH. After the compartment is full, it is sealed off to allow heat, desiccation, and alkalization to destroy faecal microbes. Sterilized waste is then spread onto household agricultural plots as fertilizer, but sterilization may be incomplete<sup>43</sup>. The village is a two-hour walk from the nearest primary health care centre, but many antibiotics are available for purchase over-the-counter. Depending on the season (rainy or dry), villagers consume a combination of food grown in individual household plots and food purchased from the town. They primarily consume beans, as well as starches such as tortillas and rice, and rarely consume meat. Chickens and cows are the most common domestic animals.

**Study design.** In the Salvadoran village (RES), the study was explained to the community in a public forum at a preliminary visit, and all members of the community were invited to participate. In the Peruvian community (PST), ten households with a minimum of four members and one child <10 years old were randomly selected and invited to participate in the study. All individuals living in the same household were invited, but were not required, to participate. Written informed consent from each participant was obtained before asking survey questions or faecal sample collection. Both studies conform to the guidelines set forth in the Helsinki Declaration. Prior to initiation, the studies were reviewed and received approval from the National Ethics Committee of El Salvador (Comité Nacional de Ética para la Investigación, Acta number 039-2012), the Institutional Review Board of Asociación Benéfica Prisma in Lima (Act CE0809.12), and The Washington University in St Louis Institutional Review Board (IRB ID numbers 201301049/201206094).

After the initial collection in January 2013, human faecal samples from RES were collected one week, three weeks, and one year later. Human faecal samples from PST were collected between January and May 2012. We sampled  $3.6 \pm 2.0$  individuals per house in RES and  $5.7 \pm 3.0$  in PST. Household environmental samples (soil, water, latrine compartments, and animal faeces from dogs, cows, chickens, guinea pigs) were collected with the permission of the residents. In RES, we sampled soils from washing areas adjacent to each house, adjacent to the latrine compartment where treated waste is removed, from urine-diverting tube exits, and where possible, from agricultural plots and chicken coops, as well as mud from the community pond's edge. We also collected water from the piped source of sand-filtered drinking water, stored drinking water, wash water from outdoor storage barrels, reservoirs, and other containers, and the community pond and its inflow.

Sewage influent (pre-treatment) and effluent (post-treatment) samples from the wastewater treatment plant 'PTAR San Juan', located in the vicinity of PST, were collected in collaboration with Water and Sewage Authority of Lima (SEDAPAL).

Twelve influent and effluent samples were collected between May 2012 and January 2013 to assess the stability of sewage communities over time. We collected influent and effluent from the districts of San Juan de Miraflores (SJM; which PST is part of) and Villa El Salvador (VES; a neighbouring pueblo joven with similar demographic characteristics), which converge in this WWTP and serve a population of ~700,000.

**RES demographic survey.** In RES, a survey was administered to each study participant at the first three sample collections inquiring about household structure, occupation, diet, and medications, among other topics. Self-reported answers were used to determine participant age, household association, and frequency of travel outside the community, as well as latrine and animal ownership for each household.

**DNA extraction.** Samples were collected in sterile containers, immediately frozen, and stored at –20 to –80 °C until shipment to Washington University in St Louis, USA where samples were stored at –80 °C until DNA extraction. Metagenomic DNA was extracted from approximately 400 to 600 mg of each faecal and latrine sample with the phenol-chloroform bead-beating protocol described previously<sup>44</sup>. Metagenomic DNA was extracted from soil and latrine samples with high ash content using MO BIO PowerSoil and PowerMax Soil DNA isolation kits. Water samples were filtered with sterile 0.22 µm filters, and metagenomic DNA was extracted from the filter membranes using the MO BIO PowerWater DNA isolation kit. For PST sewage influent and effluent, 50 ml of sample was centrifuged at 10,000 r.p.m. for 10 min; pellets were resuspended in 0.5 ml PBS, transferred to microcentrifuge tubes, and metagenomic DNA was isolated using the phenol-chloroform extraction protocol used for faecal samples. Although extraction method does have an effect in metagenome studies, large differences in community composition (such as those found between different microbial habitats) have a greater influence on variation between samples<sup>45</sup>.

**Functional metagenomic selection.** Small-insert shotgun expression libraries were created from metagenomic DNA in the vector pZE21 in *E. coli* DH10B as previously described<sup>4,44</sup>. 68 libraries were created from 51 human faecal and 45 soil, latrine, sewage, and animal faecal metagenomes. Libraries were created for all human faecal samples ( $n = 20$ ) and for soil, latrine, and cow faecal samples from two houses in RES ( $n = 14$ ). House 4 was a family of four adults, and House 6 was a family of two adults and four children where one parent routinely worked outside of the community. Libraries were also created from latrine samples from four additional houses ( $n = 4$ ). Six of the human faecal libraries, one soil library, and one animal faecal library from RES were each created from two pooled metagenomes from the same individual/location at different time points. Libraries were created from human ( $n = 31$ ) and animal ( $n = 5$ ) faecal samples from four houses in PST, as well as pooled sewage treatment plant influent and effluent. Metagenomic DNA from thirteen sewage influent and nine effluent samples were combined into a separate pool for each stage before library creation and selection because of low per-sample DNA yield.

Libraries were screened as previously described on Mueller-Hinton agar containing 50 µg ml<sup>–1</sup> kanamycin and another antibiotic at concentrations inhibitory to *E. coli* DH10B harbouring pZE21 without an insert (Supplementary Table 1). The surviving colonies for each selection were pooled. Metagenomic inserts from each pool were isolated via PCR with vector-specific primers, barcoded, and sequenced in parallel with the Illumina HiSeq 2000 (2 × 101 or 2 × 150 bp reads). Reads were demultiplexed by barcode, assembled into contigs with PARFuMS<sup>4</sup>, and annotated with Resfams v1.2 (ref. 23).

A selection was excluded from analysis of antibiotic resistance if: (a) more than 100 contigs were assembled; (b) the number of contigs assembled was more than ten times the number of colonies on the selection plate.

With these criteria, 16 out of 568 selections (2.8%) were excluded. Antibiotic resistance genes were identified by Resfam annotation (Supplementary Table 1). If a core, hand-curated, Resfam annotation specific to the antibiotic class was present on a contig, it was preferentially considered the causative resistance gene. Other Resfam annotations plausible for that antibiotic class were then identified from any contig that did not already contain a resistance gene. With this method, less-specific annotations such as efflux pumps were only identified as resistance genes if they were not co-localized with an antibiotic class-specific, canonical resistance gene.

Lipopolysaccharide modification is a conserved mechanism of antimicrobial peptide resistance in the phylum Bacteroidetes<sup>46</sup>. In the colistin selections, 25 ORFs were annotated as 'PAP2 superfamily' (PF01569.16) and shared 30.4–38.5% global amino acid identity with the *Bacteroides thetaiotaomicron* resistance gene *lpxF* (AAO76961.1)<sup>46</sup>. These open reading frames (ORFs) were also considered antibiotic resistance.

To confirm the function of the ciprofloxacin resistance gene from library 01C\_014, the plasmid was isolated from the resistant colony and reintroduced into a susceptible strain of *E. coli*, and the transformed cells were confirmed to grow in liquid and solid media containing 0.5 µg ml<sup>–1</sup> ciprofloxacin at 48 and 72 h



after inoculation, which was not observed for the negative control. The insert was amplified from the plasmid via PCR and Sanger-sequenced from both ends of the pZE21 vector. The combined Sanger sequence was manually trimmed to remove vector sequence, and the resulting contig (1,043 bp) was annotated with Resfams v1.2. The contig and its single ORF were included with the resistant contigs and ORFs identified by annotation above.

A total of 2,075 antibiotic resistance ORFs were identified on 1,955 contigs. The ORFs and the MetaGeneMark-generated protein sequences for each resistance ORF were each clustered at 100% identity over the entire length of the shorter sequence to identify identical sequences, collapsing to 1,245 unique (100% nucleotide identical) ORFs and 1,100 unique (100% amino acid identical) proteins. The contigs were clustered at 90% local identity (cd-hit-est parameters: -c 0.9 -d 0 -r 1 -G 0 -n 8 -uS 0.05 -aS 0.5) to identify different genetic surroundings.

**Identification of top hits in NCBI nr.** The MetaGeneMark-generated protein sequence for each ORF was compared to NCBI nr (accessed on 15 September 2014) with blastp to identify the top local hit(s). A Needleman-Wunsch alignment was generated with EMBOSS needle for each protein and top hit (default parameters), and the global percent identity was calculated as the number of identities over the length of the shorter sequence.

**Whole metagenome shotgun sequencing.** Metagenomic DNA was sheared to 300–400 bp, barcoded by sample, and sequenced on an Illumina HiSeq or NextSeq with 2 × 150 bp paired reads. Reads were demultiplexed by barcode with no mismatches, retaining reads whose pair did not contain the same barcode as unpaired reads. Demultiplexed reads were trimmed with Trimmomatic-0.30 to remove Illumina adaptor and low-quality bases (<Q13) from the ends, with default ILLUMINACLIP parameters and a minimum trimmed read length of 36 bases. Paired reads were trimmed in palindrome mode, while single reads were trimmed in simple mode. Human sequences were removed with DeconSeq by mapping to the human reference genome (GRCh38)<sup>47</sup>. Any paired read whose pair was a human sequence was also removed. Samples with fewer than 1 million total reads (paired and unpaired) were excluded from further analysis. 85 of the 98 samples interrogated with functional metagenomics were shotgun sequenced.

**Assembly of metagenomes from low-diversity metagenomes.** Nine shotgun metagenomes from children <3 years old with 150 OTUs/sample or fewer based on 16S data (see below) were assembled using Velvet (Supplementary Table 3). VelvetOptimiser was run on each sample with hash values from 19 to 141 in steps of 2, with both paired and unpaired reads, using n50 as the optimization function. Assembled contigs were annotated with ResFams<sup>23</sup>, and ORFs with core ResFams annotations were included in the ShortBRED markers.

**Quantification of antibiotic resistance genes in metagenomes with ShortBRED.** ShortBRED<sup>24</sup> was used to quantify the abundance of antibiotic resistance genes in the metagenomes. ShortBRED identifies unique marker sequences for clustered proteins that distinguish them from close homologues and maps reads to only those markers. This technique has greater accuracy than mapping to the entire protein, especially for antibiotic resistance genes, many of which evolved from genes performing non-resistance functions in the host.

ShortBRED markers were identified from the antibiotic resistance proteins (1) isolated from the functional selections performed in this study (2,075) (2) identified from the human faecal metagenome assemblies in this study (132), (3) the Comprehensive Antibiotic Resistance Database (CARD) (downloaded 20 October 2014; 2,972 proteins)<sup>48</sup>, and (4) the Lahey β-lactamase database (<http://www.lahey.org/studies/>; 1,145 proteins; one short protein, VEB-6, removed)<sup>49</sup> (Supplementary Table 4), clustered at 100% identity. The reference database was the modified version of the Integrated Microbial Genomes database, version 3.5., described in ref. 24. ShortBRED produced 2,275 markers when clustered at 100% identity (Supplementary Table 5) and 1,266 markers when clustered at 90% identity. Unless noted, the 100% identity markers were used for all analyses.

We quantified translated antibiotic resistance gene abundance in all metagenomes by mapping paired and unpaired fastq reads to the ShortBRED markers with 99% sequence identity. This extended our resistome investigation to individuals and sites not interrogated using functional metagenomics and to antibiotics that target Gram positive bacteria (for example, vancomycin, macrolides) and are not detectable in functional selections in our Gram negative *E. coli* host<sup>50</sup>. All analyses were performed on marker abundances normalized to reads per kilobase per million reads (RPKM).

For antibiotic resistance proteins from the CARD and Lahey databases, metadata (resistance category, mechanism of action, antibiotic target(s)) was hand-curated from information available on the CARD website. For antibiotic resistance proteins identified through functional metagenomic selections and shotgun assemblies, resistance category and mechanism of action were assigned based on Resfams annotation. Antibiotic target(s) for the former were the antibiotics to which they conferred resistance in our functional selections, while the latter were

not assigned an antibiotic target. Annotations for ShortBRED markers were drawn from all constituent proteins.

**Comparison of human faecal resistomes to published cohorts.** For comparison to the resistomes in ref. 25, whole metagenome shotgun reads were downloaded from SRA (accession PRJNA268964). Fastq reads were trimmed with Trimmomatic in simple mode using the same parameters as for the RES and PST reads. Samples with fewer than 1 million total reads and individuals <3 years old were excluded. ShortBRED markers were quantified as above. For ref. 25, the average read length of the paired reads for each sample was specified as the average read length during ShortBRED quantification. Only ShortBRED markers that included proteins from the CARD and Lahey databases were considered for this analysis to avoid bias towards our cohorts.

**16S gene V4 amplification, sequencing, and preprocessing.** The 16S gene V4 region (515–806) was amplified using the original Earth Microbiome Project (EMP) protocol (<http://www.earthmicrobiome.org/emp-standard-protocols/16s/>) or with Takara Taq DNA polymerase premix, with barcoded primers designed in ref. 51, from 228 human faecal and 203 environmental samples. 245 samples interrogated by whole metagenome shotgun sequencing were also 16S sequenced. Barcoded amplicons were pooled and sequenced on an Illumina MiSeq with 2 × 250 bp paired-end reads.

Barcoded Illumina reads were demultiplexed with QIIME version 1.8.0, *split\_libraries\_fastq.py*<sup>52</sup>. Paired reads were truncated at the first base with quality score ≤ Q3 and merged using usearch<sup>53</sup>, requiring 100% identity in the overlap region and a merged length of 253 bp ± 5 bp. Merged reads were filtered with QIIME to remove reads with three or more contiguous bases with quality score ≤ Q20.

**Open OTU picking with UPARSE for RES and PST comparisons.** OTUs were picked from all 12,797,788 merged and filtered RES and PST reads with the UPARSE pipeline<sup>54</sup>. Singletons were excluded from OTU picking, and a reference-based chimera check against the GOLD database (downloaded 1 September 2013) was performed on OTUs as recommended. Reads were assigned to OTUs at 97% identity using usearch. Representative sequences from each OTU were assigned taxonomy with uclust against the Greengenes database (release 13\_8, 97% clusters), aligned, and used to create a phylogenetic tree using QIIME. 19,301 OTUs were picked across all samples. Biom tables were rarefied to 7,000 sequences per sample, which excluded eight samples.

**Comparison of phylogenetic composition of human faecal samples to published cohorts.** Raw 16S amplicons from ref. 19 were downloaded from MG-RAST (accession number qiime:850). 16S amplicons from faecal samples from ref. 14 were downloaded from the European Nucleotide Archive (ENA) (ERP008799). 16S amplicons from ref. 25 were provided by the authors. For all studies, reads were generated with primers F515/R806 on an Illumina platform. We classified each cohort's lifestyle as described in ref. 25. Only samples from individuals 3 years or older were included in analysis. Individuals labelled with an adult-specific keyword (Family Member: "Mother", "Father", or Sample Identifier: "adlt") in ref. 19 supplementary table 2 were also included. Samples from ref. 19 were rarefied to 50,000 reads each to reduce computational load. The merged reads from RES and PST human samples and the reads from ref. 14 were truncated at 101 bp.

OTUs were picked from all reads with the reference-based protocol described in ref. 19 against the Greengenes database (13\_8, 97% clusters). Reverse strand alignment was permitted, and new clusters were suppressed. The OTU table was rarefied to 5,000 sequences per sample, which excluded two samples from ref. 14.

**Sample filters.** To better characterize within- and between-sample diversity, we limited both the shotgun and 16S analysis to a single sample per human individual except where noted. Infant faecal microbiota undergo large and frequent shifts in microbial composition before stabilizing into an adult-like configuration around three years of age<sup>19</sup>. Therefore, except where noted, we limited all analyses with human samples to those from individuals 3 years old or older.

**Alpha diversity.** For 16S, biom tables were rarefied ten times to 7,000 sequences per sample, and equitability, observed species, and Faith's phylogenetic distance were calculated on all rarefactions and averaged using the QIIME script *alpha\_diversity.py*. For resistome, tables were not rarefied, and only the number of ARGs present in each sample was compared. Alpha diversity was compared between categories using non-parametric Student's *t*-tests with 999 permutations at a depth of 7,000 sequences per sample with Bonferroni correction for multiple hypothesis testing using the QIIME script *compare\_alpha\_diversity.py*. For Extended Data Fig. 4g, the total antibiotic resistance RPKM for each sample was summed, multiplied by one million, and divided by the total number of reads for that sample to confirm that the results were robust to sequencing depth.

**Beta diversity.** Unweighted UniFrac, weighted UniFrac, Sørensen-Dice, and Bray-Curtis dissimilarity matrices were calculated from biom tables with the QIIME script *beta\_diversity.py*, using the Greengenes 13\_8 97% phylogenetic tree (for closed reference OTU picking) and the phylogenetic tree generated during

*de novo* OTU picking for UniFrac distances. Principal coordinates analysis was performed with the QIIME script *principal\_coordinates.py*.

ANOSIM and adonis analyses were performed on dissimilarity matrices with the QIIME script *compare\_categories.py* or in R (vegan package functions *anosim()* and *adonis()*) with 999 permutations.

Distance to centroid was calculated in R with the vegan package function *betadist()*, and significance was tested with a permutation test (*permutest()*).

Average pairwise dissimilarities between categories were compared with non-parametric Student's *t*-tests with 999 Monte Carlo permutations with Bonferroni correction using code adapted from the QIIME script *make\_distance\_boxplots.py*. For shared antibiotic resistance proteins, the number of antibiotic resistance proteins shared between each pair of samples was calculated with *shared\_phylotypes.py*, and significance was calculated as above.

**Procrustes.** The 16S OTU table rarefied to 7,000 sequences/sample and the antibiotic resistance abundance table were both filtered to include only samples interrogated with both methods ( $n = 172$ ). Bray–Curtis dissimilarity matrices were calculated for both, principal coordinates analysis was performed, and procrustes analysis was performed on the PCoA results (QIIME script *transform\_coordinate\_matrices.py*) across 172 dimensions with 999 permutations to determine significance.

**Identification of discriminatory taxa using LEfSe.** Taxa summaries were created from the rarefied OTU table and filtered to levels L1 to L5 (kingdom – family). LEfSe<sup>55</sup> was used to identify taxa that were discriminative between categories. Alpha was 0.05 for both the Kruskal–Wallis and pairwise Wilcoxon rank-sum tests. The Linear Discriminant Analysis (LDA) effect size threshold was set at 3.0 or 4.0, and all-against-all comparison was performed.

**Identification of discriminatory antibiotic resistance functions.** The antibiotic resistance table was filtered to only antibiotic resistance proteins detected in the sample set under consideration and summarized by each metadata category with the QIIME script *summarize\_taxa.py*. Absolute abundances rather than relative abundances were calculated, because antibiotic resistance genes represent a small fraction of the total metagenome. Supervised learning with the Random Forests classifier was performed to identify antibiotic resistance categories that most discriminated between sample categories (*supervised\_learning.py*, 500 trees, and tenfold cross-validation). The top discriminatory antibiotic resistance categories (as determined by feature importance scores) enriched in subsets of the samples were plotted onto the principal coordinates of all samples, using the abundance-weighted average for the coordinates. The size of the point is proportional to the overall abundance across all samples under consideration (1/100 RPKM).

**SourceTracker.** We estimated the proportion of each latrine, soil, and water microbiota attributable to RES human faeces and the proportion of each sewage microbiota attributable to PST human faeces using SourceTracker<sup>29</sup>. Antibiotic resistance tables were filtered to remove any samples and markers without any observations. SourceTracker was run through QIIME with default settings using human faecal microbiota as source. Jigger was introduced in boxplots to allow visualization, but some samples with very low abundance are not distinguishable.

**Mass spectrometry-based detection of antibiotics in WWTP samples.** Solid phase extraction and ultra-performance liquid chromatography–tandem mass spectrometry (UPLC–MS/MS) were used to detect the presence of 16 antibiotics from 7 classes in 11 influent and 11 effluent sample supernatants using a modified protocol<sup>33</sup>. Fifty ml of influent or effluent sample were centrifuged at 10,000 r.p.m. for 10 min; supernatants were concentrated using solid phase extraction cartridges (6cc, 200 mg, Waters, Milford, MA). Samples were analysed on a 4000 QTRAP LC/MS/MS instrument (AB Sciex). The data was normalized based on the internal standard (isotopically labelled caffeine, 13C3, Cambridge Isotope Laboratories) to account for experimental variation and antibiotic extraction/ionization efficiency. A mixture of the antibiotic standards was also processed along with the samples as positive controls to test for recovery. Standards for amoxicillin, cefotaxime, cefoxitin, chloramphenicol, ciprofloxacin, clindamycin, erythromycin, penicillin G, sulfamethoxazole, tetracycline, and trimethoprim were purchased from Sigma-Aldrich (St Louis, MO). Standards for azithromycin, aztreonam, cefepime, doxycycline, and tigecycline were purchased from AK Scientific (Union City, CA). All analyses were performed at the Proteomics and Mass Spectrometry Facility at the Donald Danforth Plant Science Center (St Louis, MO).

**Network creation.** Antibiotic resistance gene networks were generated based on ShortBRED results of human and environmental metagenomes, filtering hits with

RPKM value of <0.1 to determine sharing across samples. Graphics were generated in Cytoscape 3.2.1 using an edge-weighted spring-embedded layout.

**Mobile genetic elements (MGEs).** Putative MGEs were identified from the functional selection contigs by Pfam and TIGRfam annotation (Supplementary Table 13). Antibiotic-resistant ORFs were considered co-localized with an MGE if they shared a contig with a MGE ORF. 365 (19%) antibiotic resistance contigs contained at least one of 236 unique MGEs (100% amino acid identical).

**Multidrug resistance clusters.** Multidrug resistance clusters were identified as contigs containing multiple antibiotic resistance proteins identified from our selections and/or annotated with a core Resfams profile HMM (ref. 23), regardless of the antibiotic used in the selection. 200 (10%) contigs contained multidrug resistance clusters.

**Mobilome analysis.** For ‘mobilome’ analyses, the ShortBRED markers were restricted to those that included an antibiotic resistance protein encoded in our functional metagenomic selection contigs adjacent to a putative MGE.

**Phylogenetic classification of contigs with PhyloPythia.** Antibiotic-resistant contigs and metagenomic assembly contigs were classified by PhyloPythia<sup>56,57</sup> using the recommended model (800 genera, 2013).

**Data reporting.** No statistical methods were used to predetermine sample size. The experiments were not randomized. The investigators were not blinded to allocation during experiments and outcome assessment.

37. Checkley, W. *et al.* Effect of water and sanitation on childhood health in a poor Peruvian peri-urban community. *Lancet* **363**, 112–118 (2004).
38. Cooper, M. A. *et al.* Molecular analysis of household transmission of *Giardia lamblia* in a region of high endemicity in Peru. *J. Infect. Dis.* **202**, 1713–1721 (2010).
39. Oswald, W. E. *et al.* Fecal contamination of drinking water within peri-urban households, Lima, Peru. *Am. J. Trop. Med. Hyg.* **77**, 699–704 (2007).
40. Water Sanitation and Health Unit Organization ([http://www.who.int/water\\_sanitation\\_health/sanitproblems/en/index4.html](http://www.who.int/water_sanitation_health/sanitproblems/en/index4.html)) (World Health Organization, 2002).
41. WHO / UNICEF Joint Monitoring Programme (JMP) for Water Supply and Sanitation. *El Salvador: estimates on the use of water sources and sanitation facilities (1980–2015)*. (2015).
42. The World Bank Group. *Data: Topics: Urban Development* (<http://data.worldbank.org/topic/urban-development>) (2015).
43. Corrales, L. F., Izurieta, R. & Moe, C. L. Association between intestinal parasitic infections and type of sanitation system in rural El Salvador. *Trop. Med. & Int. Health* **11**, 1821–1831 (2006).
44. Moore, A. M. *et al.* Pediatric fecal microbiota harbor diverse and novel antibiotic resistance genes. *PLoS ONE* **8**, e78822 (2013).
45. Lozupone, C. A. *et al.* Meta-analyses of studies of the human microbiota. *Genome Res.* **23**, 1704–1714 (2013).
46. Cullen, T. W. *et al.* Antimicrobial peptide resistance mediates resilience of prominent gut commensals during inflammation. *Science* **347**, 170–175 (2015).
47. Schmieder, R. & Edwards, R. Fast identification and removal of sequence contamination from genomic and metagenomic datasets. *PLoS ONE* **6**, e17288 (2011).
48. McArthur, A. G. *et al.* The comprehensive antibiotic resistance database. *Antimicrob. Agents Chemother.* **57**, 3348–3357 (2013).
49. Bush, K., Palzkill, T. & Jacoby, G. <http://www.lahey.org/studies/> (Lahey Clinic, 2015).
50. Pehrsson, E. C., Forsberg, K. J., Gibson, M. K., Ahmadi, S. & Dantas, G. Novel resistance functions uncovered using functional metagenomic investigations of resistance reservoirs. *Front. Microbiol.* **4**, 145 (2013).
51. Caporaso, J. G. *et al.* Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms. *ISME J.* **6**, 1621–1624 (2012).
52. Caporaso, J. G. *et al.* QIIME allows analysis of high-throughput community sequencing data. *Nature Methods* **7**, 335–336 (2010).
53. Edgar, R. C. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **26**, 2460–2461 (2010).
54. Edgar, R. C. UPARSE: highly accurate OTU sequences from microbial amplicon reads. *Nature Methods* **10**, 996–998 (2013).
55. Segata, N. *et al.* Metagenomic biomarker discovery and explanation. *Genome Biol.* **12**, R60 (2011).
56. McHardy, A. C., Martin, H. G., Tsirigos, A., Hugenholtz, P. & Rigoutsos, I. Accurate phylogenetic classification of variable-length DNA fragments. *Nature Methods* **4**, 63–72 (2007).
57. Patil, K. R., Roune, L. & McHardy, A. C. The PhyloPythiaS web server for taxonomic assignment of metagenome sequences. *PLoS ONE* **7**, e38581 (2012).

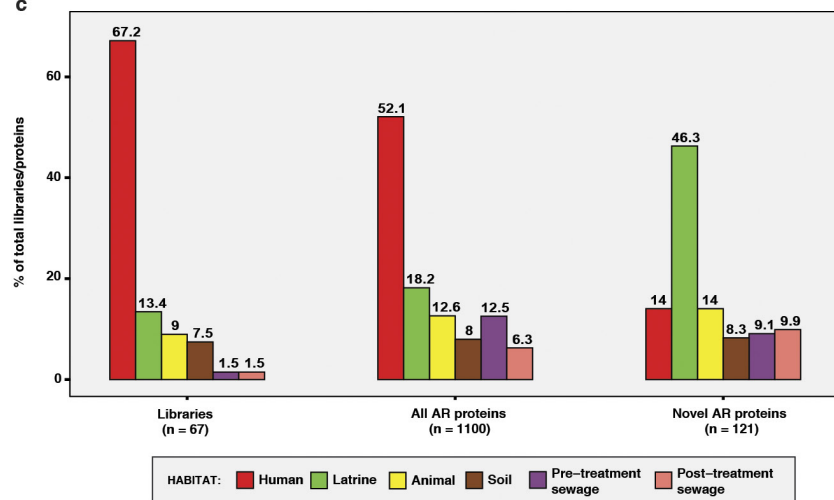
a



b

	Samples interrogated with 16S seq.	OTUs	Samples interrogated with shotgun seq.	AR genes
Human	226	1641	191	584
Post-treatment sewage	13	2220	7	82
Pre-treatment sewage	30	2324	27	514
Animal	14	2409	10	283
Water	22	3025	4	8
Latrine	36	3245	16	242
Soil	84	12503	30	206
Total	425	19301	285	797

c

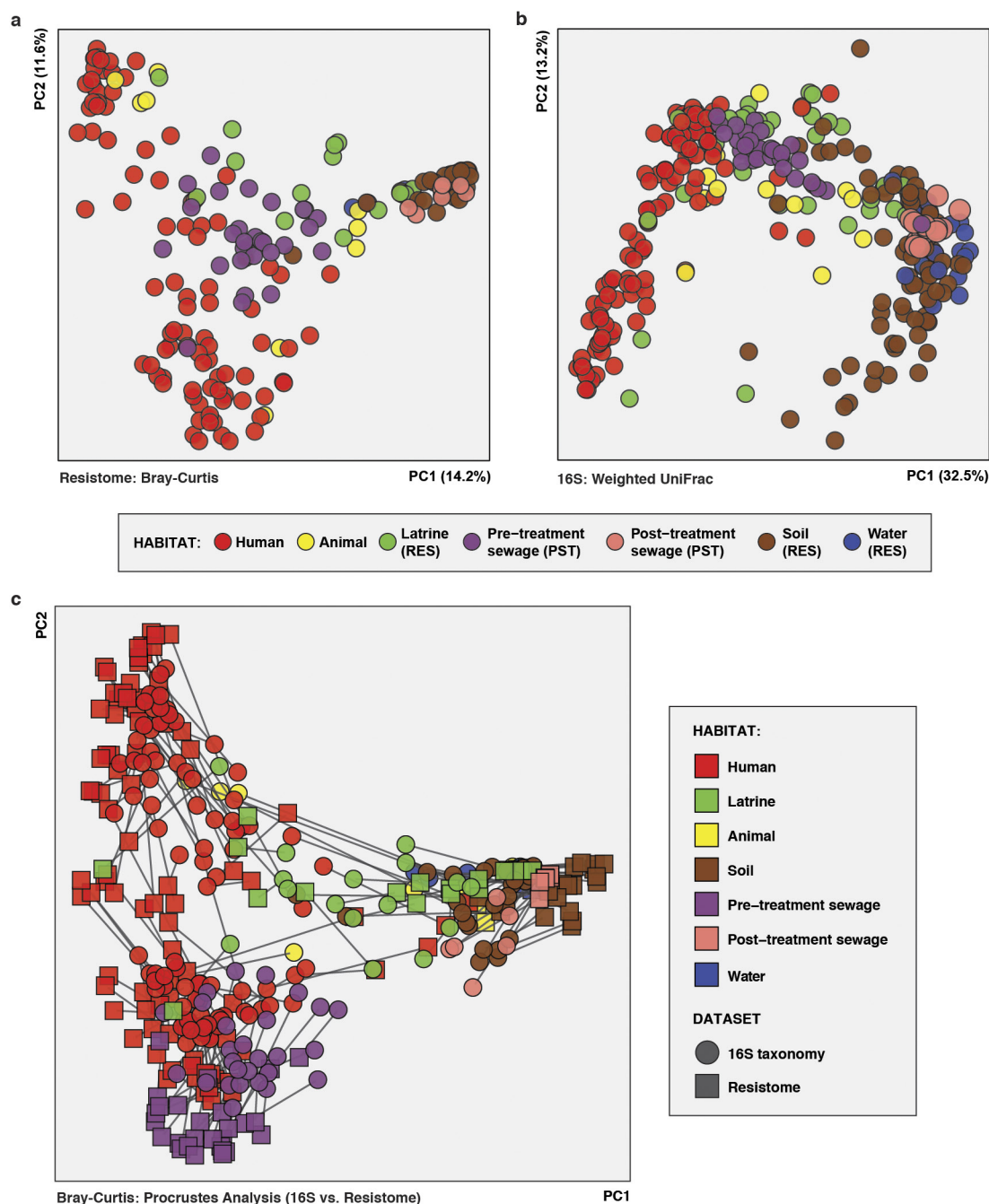


**Extended Data Figure 1 | Overview of study and methods.** a, Location and overview of study sites in El Salvador and Peru. RES photographs by the authors G.S.S. and M.M.B., PST photographs by the author P.T.

b, Antibiotic resistance markers and OTUs detected vs number of samples interrogated by whole metagenome and 16S sequencing by habitat in RES and PST. c, Proportion of metagenomic libraries ( $n = 67$ ), all antibiotic resistance proteins identified from functional metagenomic selections ( $n = 1,100$ ), and novel antibiotic resistance proteins identified from

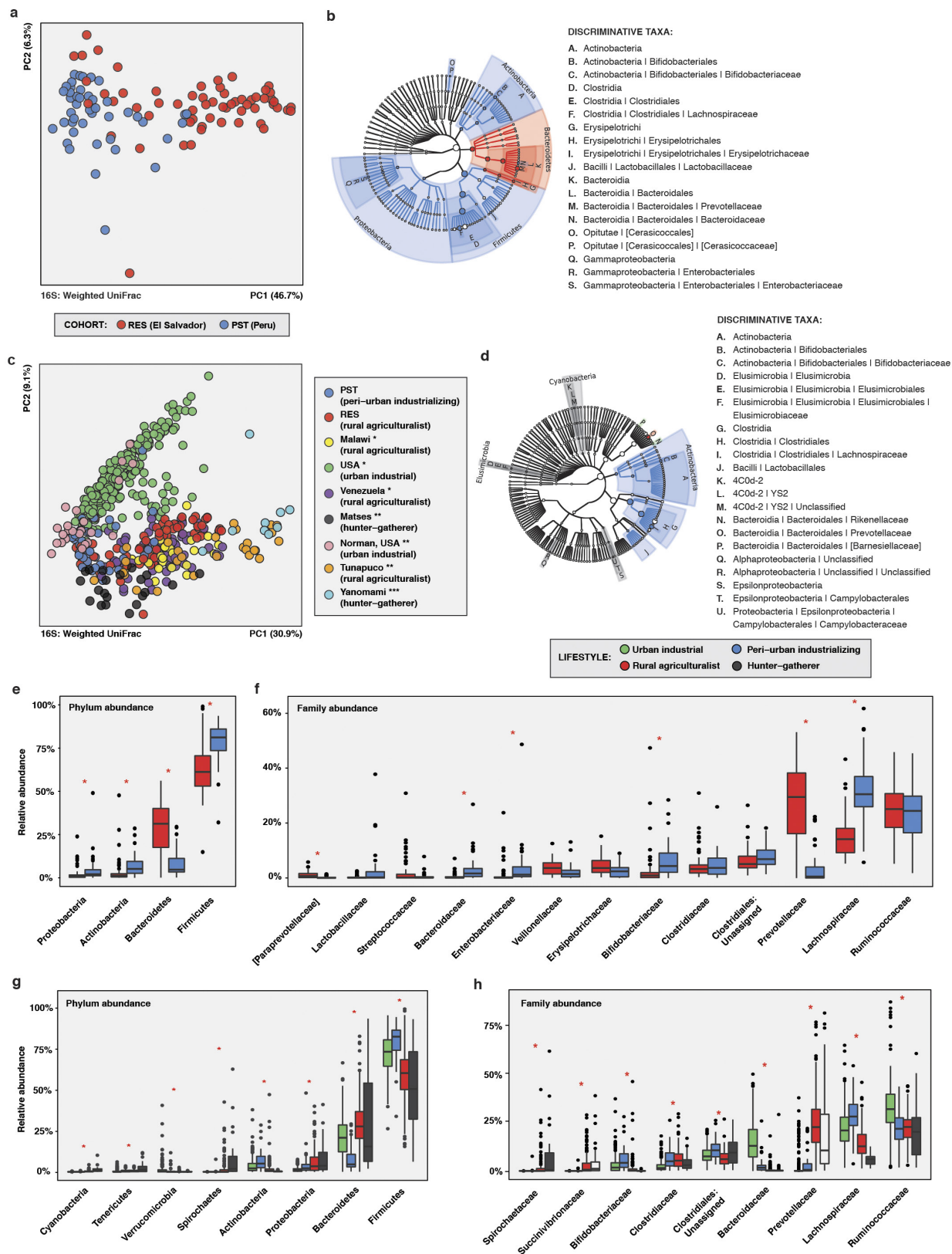
functional metagenomic selections ( $n = 121$ ) originating from each microbial habitat. The percent of total libraries/proteins in that category originating from each microbial habitat is listed above the bar. For all antibiotic resistance proteins and novel antibiotic resistance proteins, the total sums to  $>100\%$  due to proteins identified in more than one habitat. The number of novel antibiotic resistance proteins vs libraries screened was significantly different than expected compared to the total for human and latrines (chi-squared test,  $P < 0.005$ ).





**Extended Data Figure 2 | Human faecal and environmental microbiota from RES and PST.** Microbiota are coloured by habitat. **a**, PCoA of Bray-Curtis distances between resistomes. ( $n = 86$ ,  $n = 10$ ,  $n = 16$ ,  $n = 30$ ,  $n = 4$ ,  $n = 27$  and  $n = 7$  for human, animal, latrine, soil, water, pre-treatment sewage and post-treatment sewage, respectively) Adonis  $R^2 = 22.4\%$ ,  $P < 0.001$ . **b**, PCoA of weighted UniFrac distances between

microbiota. ( $n = 105$ ,  $n = 14$ ,  $n = 36$ ,  $n = 84$ ,  $n = 22$ ,  $n = 30$  and  $n = 13$  for human, animal, latrine, soil, water, pre-treatment sewage and post-treatment sewage, respectively.) Adonis  $R^2 = 41.9\%$ ,  $P < 0.001$ . **c**, Procrustes transformation of taxonomic composition vs resistome. Only samples interrogated with both methods were included ( $n = 172$ ).  $M^2 = 0.360$ ,  $P < 0.001$  (172 dimensions, 999 permutations).

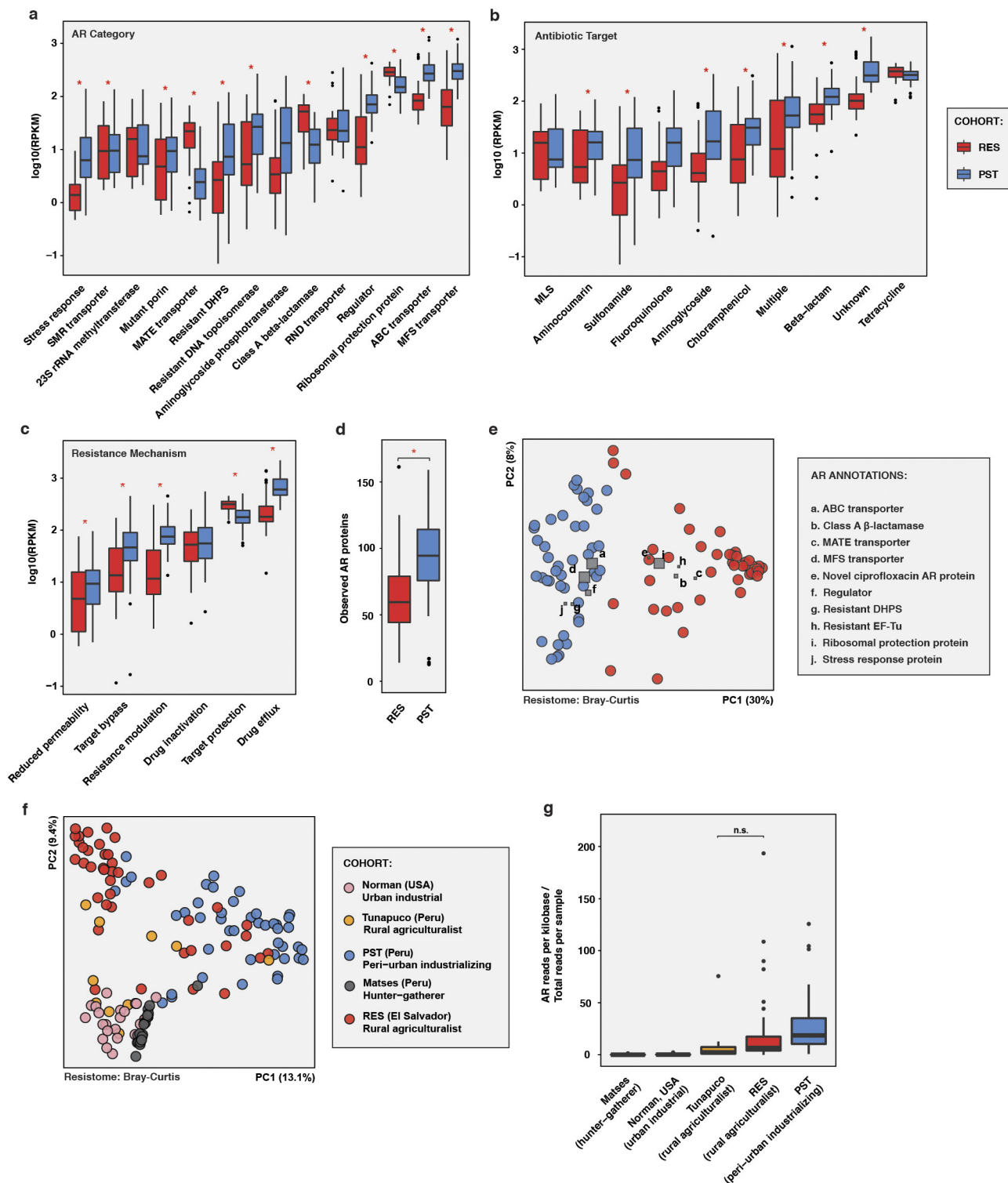


Extended Data Figure 3 | See next page for caption.

**Extended Data Figure 3 | Phylogenetic composition of RES and PST human faecal microbiota and published microbiota from previous studies<sup>14,19,25</sup>.** **a, b, e, f**, RES vs PST. (RES  $n = 60$ , PST  $n = 45$ ) **c, d, g, h**, RES and PST vs published human microbiota. (RES  $n = 60$ , PST  $n = 46$ , other  $n = 446$ ; see Supplementary Table 14) **a**, PCoA of weighted UniFrac distances between RES and PST human faecal microbiota, coloured by cohort. Adonis  $R^2 = 29.7\%$ ,  $P < 0.001$ . **b**, Taxa discriminating between RES and PST human faecal microbiota as determined by LEfSe. The phylogenetic tree includes all kingdom- to family-level taxa present in any sample. Coloured taxa are discriminative between cohorts and have an LDA effect size of  $\geq 4.0$ ; they are coloured by the cohort in which they have the highest abundance. Circle size is relative to the highest abundance in either cohort. **c**, PCoA of weighted UniFrac distances between RES and PST human faecal microbiota and published human faecal microbiota,

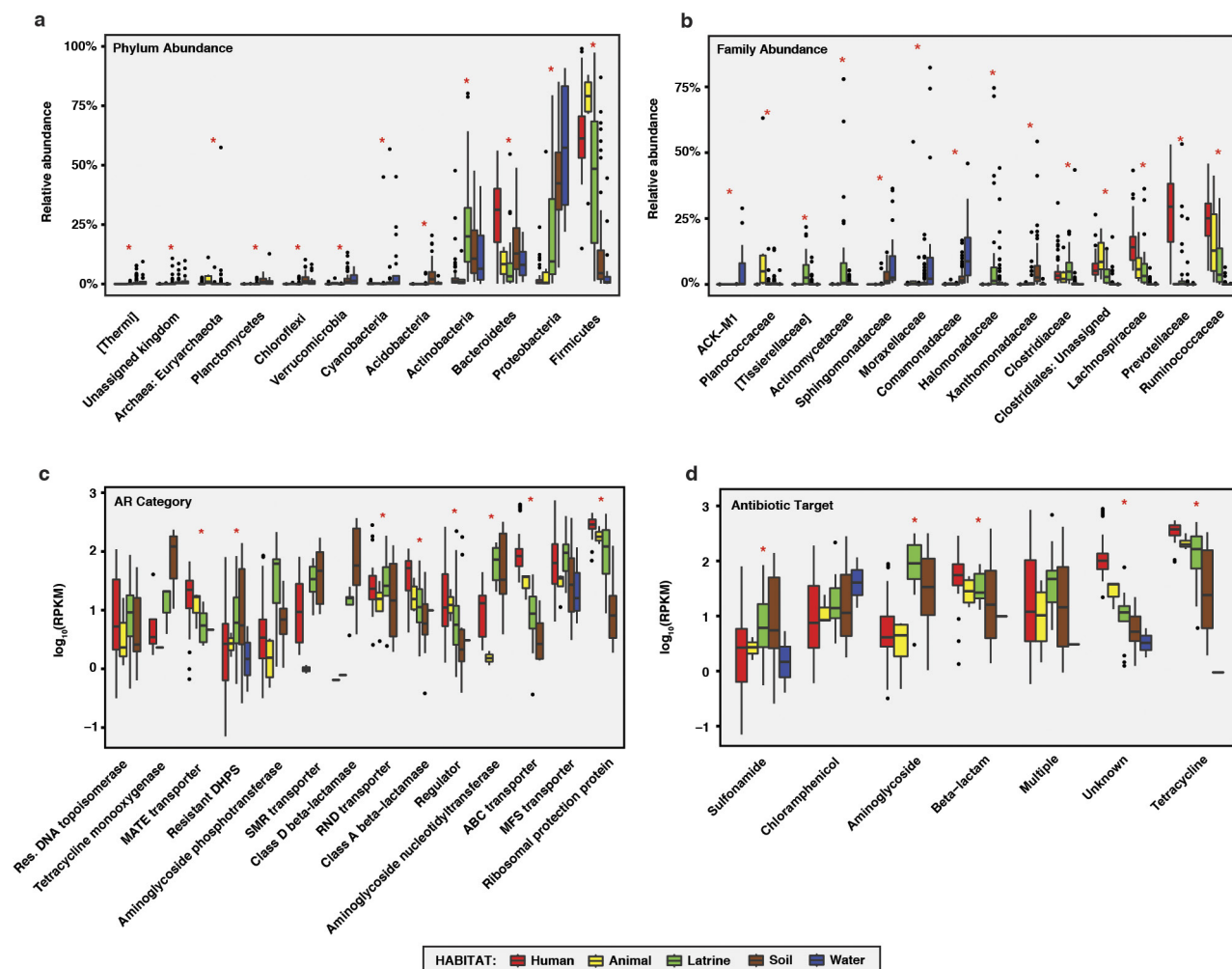
coloured by cohort. Cohorts are labelled by lifestyle and study (\* (ref. 19), \*\* (ref. 35), \*\*\* (ref. 14)). Adonis  $R^2 = 37.6\%$ ,  $P < 0.001$ . **d**, Taxa discriminating between host lifestyles for RES and PST and published human faecal microbiota as determined by LEfSe, effect size threshold 3.0. Discriminative taxa are coloured by the host lifestyle in which they are most abundant. **e, f**, Relative abundances of microbial phyla (**e**) and families (**f**) in human faecal microbiota from RES and PST. \* $P < 0.05$ , Wilcoxon test with Bonferroni correction. **g, h**, Relative abundances of microbial phyla (**g**) and families (**h**) in human faecal microbiota from RES and PST and published human faecal microbiota, by lifestyle. \* $P < 0.05$ , Kruskal–Wallis test with Bonferroni correction. **e–h**, Only taxa with a mean relative abundance of  $\geq 1\%$  in one cohort/lifestyle are shown. Taxa are in order of increasing overall mean relative abundance. Error bars, s.d.; centre bars, median.





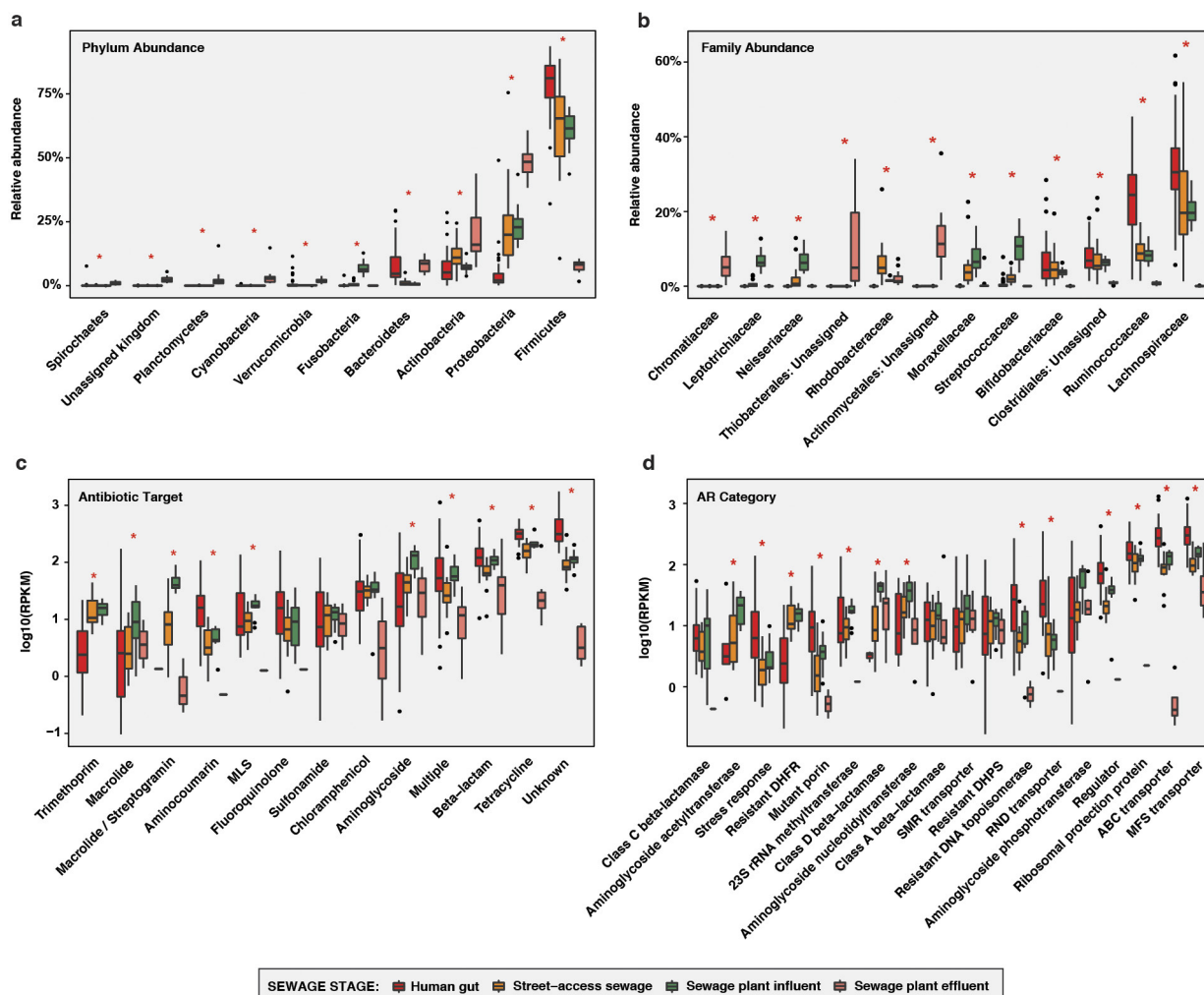
**Extended Data Figure 4 | RES and PST human faecal resistomes and comparison to the published data sets from ref. 25.** **a–e**, RES and PST resistomes, coloured by cohort. (RES  $n = 42$ , PST  $n = 44$ ) **f**, **g**, RES and PST vs published human data sets, coloured by cohort. (RES  $n = 42$ , PST  $n = 44$ , other  $n = 53$ ; see Supplementary Table 15) **a–c**, Absolute abundances of antibiotic resistance categories (**a**), antibiotic targets (**b**), and mechanisms of action (**c**) in human faecal resistomes from RES and PST. Only categories with a mean RPKM of  $>10$  in one cohort are shown. Categories are in increasing order of overall mean absolute abundance. Abundances are plotted in  $\log_{10}$  scale.  $*P < 0.05$ , Wilcoxon test with Bonferroni correction. **d**, Number of antibiotic resistance proteins per RES and PST human faecal resistome.  $*P < 0.05$ , non-parametric Student's  $t$ -tests. **e**, PCoA of Bray–Curtis distances between RES and PST resistomes,

with abundance-weighted coordinates of the top five most discriminative antibiotic resistance categories enriched in each cohort (squares, size proportional to overall abundance). Adonis  $R^2 = 25.0\%$ ,  $P < 0.001$ . **f**, PCoA of Bray–Curtis distances between human faecal resistomes from RES and PST and ref. 25. Adonis  $R^2 = 19.7\%$ ,  $P < 0.001$ . **g**, Total reads mapping to antibiotic resistance markers per person (normalized by marker length) normalized by the total reads in that sample in RES and PST and published human faecal microbiota, by cohort. Includes both paired and unpaired reads. The overall distribution of normalized antibiotic resistance read depth was significantly different than expected (Kruskal–Wallis,  $P < 1 \times 10^{-15}$ ). n.s., not significant. All other comparisons are  $P < 0.05$ , Wilcoxon test with Bonferroni correction. **a–d**, **f**, Error bars, s.d.; centre bars, median.



**Extended Data Figure 5 | RES human faecal and environmental microbiota and resistomes.** **a, b**, Relative abundances of microbial phyla (**a**) and families (**b**) in RES microbiota, by habitat. ( $n = 60$ ,  $n = 6$ ,  $n = 36$ ,  $n = 84$  and  $n = 22$  for human, animal, latrine, soil, water, respectively) Only taxa with a mean relative abundance of  $\geq 1\%$  in one habitat are shown. Taxa are in increasing order of overall mean relative abundance.  $*P < 0.05$ , Kruskal–Wallis test with Bonferroni correction. **c, d**, Absolute

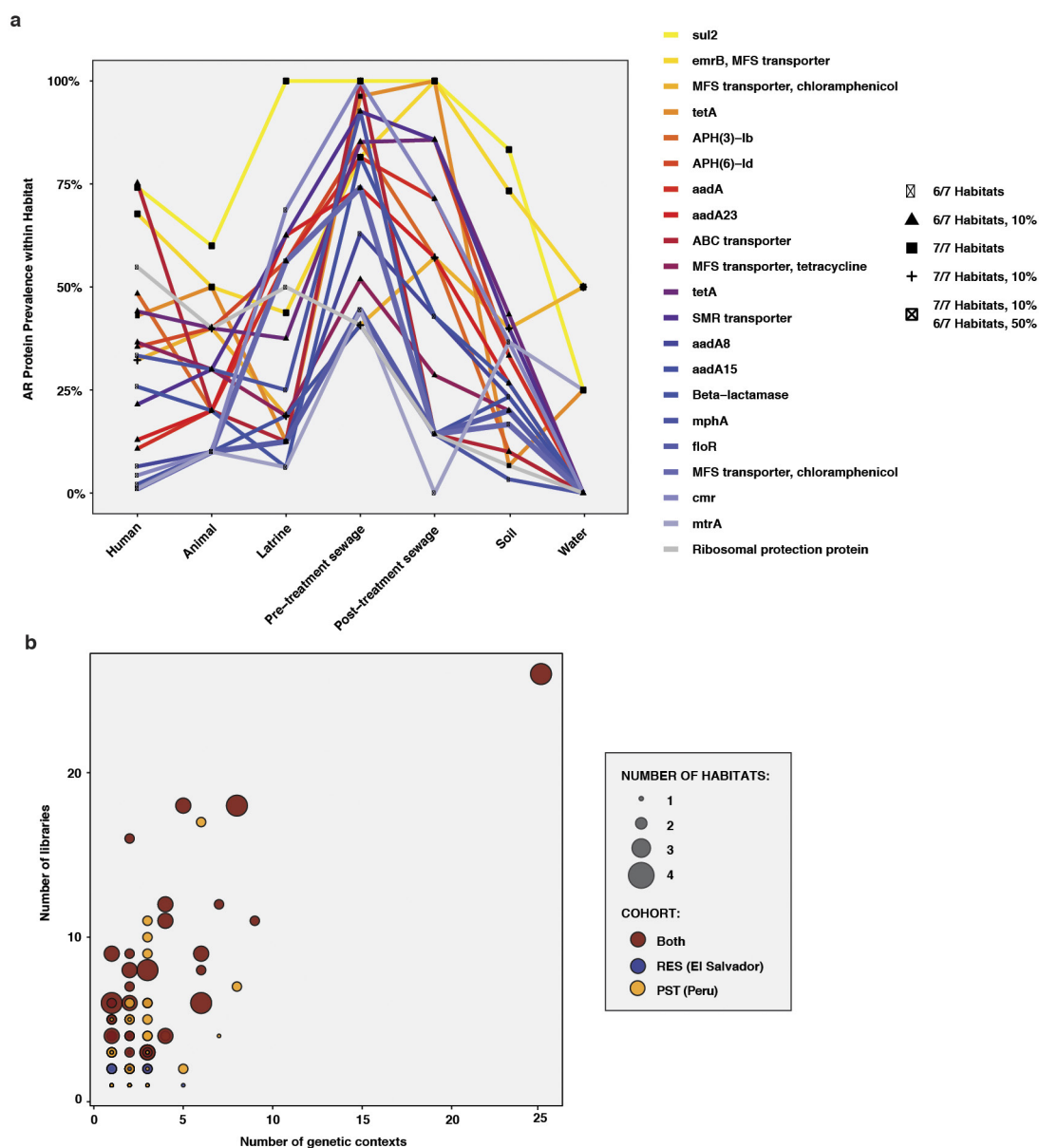
abundances of antibiotic resistance categories (**c**) and antibiotic targets (**d**) in RES resistomes, by habitat. ( $n = 42$ ,  $n = 4$ ,  $n = 16$ ,  $n = 30$  and  $n = 4$  for human, animal, latrine, soil, water, respectively). Only categories with a mean RPKM of  $> 10$  in one habitat are shown. Categories are in increasing order of overall mean absolute abundance. Abundances are plotted in  $\log_{10}$  scale.  $*P < 0.05$ , Kruskal–Wallis test with Bonferroni correction. **a–d**, Error bars, s.d.; centre bars, median.



**Extended Data Figure 6 | PST human faecal and environmental microbiota and resistomes.** **a**, **b**, Relative abundances of microbial phyla (**a**) and families (**b**) in human faecal and sewage microbiota from PST, by stage. ( $n = 45$ ,  $n = 16$ ,  $n = 14$  and  $n = 13$  for human, street-access, influent and effluent, respectively) Only taxa with a mean relative abundance of  $\geq 1\%$  in one stage are shown. Taxa are in increasing order of overall mean relative abundance.  $*P < 0.05$ , Kruskal–Wallis test with Bonferroni correction. **c**, **d**, Absolute abundances of antibiotic resistance categories

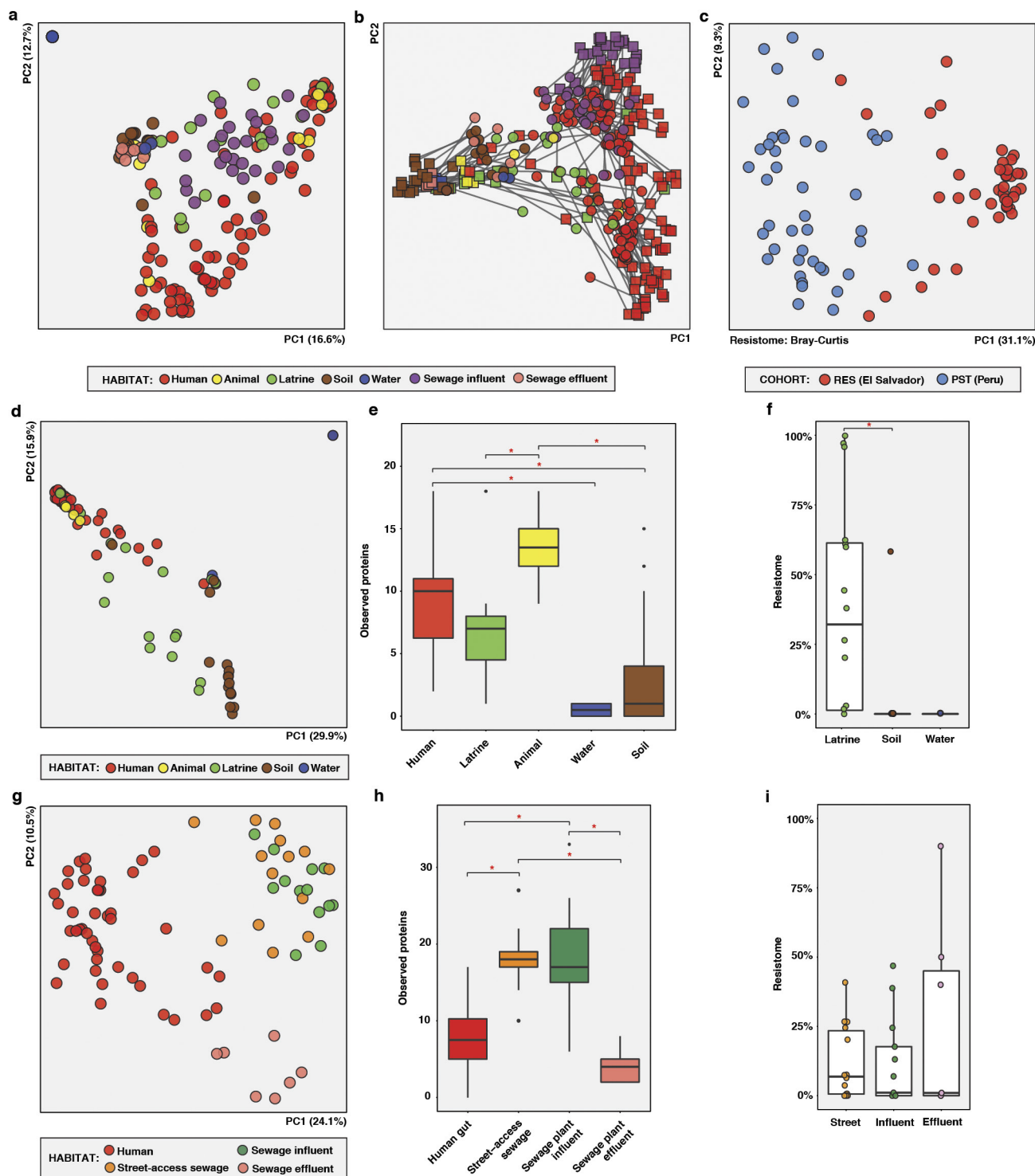
(**c**) and antibiotic targets (**d**) in PST resistomes, by stage. ( $n = 44$ ,  $n = 14$ ,  $n = 13$  and  $n = 7$  for human, street-access, influent and effluent, respectively). Only categories with a mean RPKM of  $> 10$  in one stage are shown. Categories are in increasing order of overall mean absolute abundance. Abundances are plotted in  $\log_{10}$  scale.  $*P < 0.05$ , Kruskal–Wallis test with Bonferroni correction. **a–d**, Error bars, s.d.; centre bars, median.





**Extended Data Figure 7 | Antibiotic resistance gene sharing across habitats.** **a**, Highly cosmopolitan antibiotic resistance proteins. The prevalence of each antibiotic resistance protein in metagenomes from each microbial habitat is depicted for all proteins detected in six of the seven habitats ( $n = 21$ ). Detection was based on ShortBRED quantification of the protein in each metagenome. Prevalences for an antibiotic resistance protein are linked by lines of the same colour. The shape of each point reflects the number of habitats in which it was found, as well as the minimum prevalence within each habitat. The legend lists the annotation for each protein. **b**, Protein sequences of antibiotic resistance genes isolated from functional metagenomic selections were clustered

at 100% amino acid identity, and the number of metagenomic libraries, microbial habitats (for example, human faecal, soil), and cohorts in which each unique protein ( $n = 1,100$ ) was encoded were calculated across all members of the cluster. Antibiotic resistance contigs ( $n = 1,955$ ) were clustered at 90% local identity to identify different genetic contexts, and the number of genetic contexts in which each unique protein was encoded was calculated across all contigs encoding a protein in that cluster. Spearman's  $\rho = 0.59$ ,  $P < 2.2 \times 10^{-16}$ , number of genetic contexts vs libraries;  $\rho = 0.47$ ,  $P < 2.2 \times 10^{-16}$ , number of genetic contexts vs. habitats; Wilcoxon test,  $P < 2.2 \times 10^{-16}$ , number of genetic contexts vs cohorts (one or both).



**Extended Data Figure 8 | Mobilome analyses.** **a**, PCoA of Bray-Curtis distances between RES and PST human and environmental resistomes, coloured by habitat. (n = 86, n = 10, n = 16, n = 30, n = 4, n = 27 and n = 7 for human, animal, latrine, soil, water, pre-treatment sewage and post-treatment sewage, respectively) Adonis  $R^2 = 24.1\%$ ,  $P < 0.001$ . **b**, Procrustes transformation of taxonomic composition vs resistome. Only samples interrogated with both methods were included (n = 172).  $M^2 = 0.493$ ,  $P < 0.001$  (172 dimensions, 999 permutations). **c**, PCoA of Bray-Curtis distances between RES (n = 42) and PST (n = 44) resistomes, coloured by cohort. Adonis  $R^2 = 31.0\%$ ,  $P < 0.001$ . **d-f**, RES human faecal and environmental microbiota and resistomes, coloured by habitat. (n = 42, n = 4, n = 16, n = 30 and n = 4 for human, animal, latrine, soil, water, respectively). **d**, PCoA of Bray-Curtis distances between resistomes.

Adonis  $R^2 = 32.0\%$ ,  $P < 0.001$ . **e**, Observed antibiotic resistance proteins. \* $P < 0.05$ , non-parametric Student's *t*-tests, Bonferroni correction. **f**, Percentage of latrine, soil, and water resistomes attributable to human faeces, as determined by SourceTracker<sup>29</sup>. \* $P < 0.05$ , pairwise Wilcoxon tests, Bonferroni correction. **g-i**, PST human faecal and sewage microbiota and resistomes, coloured by stage. (n = 44, n = 14, n = 13 and n = 7 for human, street-access, influent, effluent, respectively). **g**, PCoA of Bray-Curtis distances between resistomes. Adonis  $R^2 = 34.8\%$ ,  $P < 0.001$ . **h**, Observed antibiotic resistance proteins. \* $P < 0.05$ , non-parametric Student's *t*-tests, Bonferroni correction. **i**, Percentage of sewage resistomes attributable to human faeces at each sewage treatment stage, as determined by SourceTracker. \* $P < 0.05$ , pairwise Wilcoxon tests, Bonferroni correction. Error bars, s.d.; centre bars, median.

# No Sun-like dynamo on the active star $\zeta$ Andromedae from starspot asymmetry

R. M. Roettenbacher<sup>1</sup>, J. D. Monnier<sup>1</sup>, H. Korhonen<sup>2,3</sup>, A. N. Aarnio<sup>1</sup>, F. Baron<sup>1,4</sup>, X. Che<sup>1</sup>, R. O. Harmon<sup>5</sup>, Zs. Kővári<sup>6</sup>, S. Kraus<sup>1,7</sup>, G. H. Schaefer<sup>8</sup>, G. Torres<sup>9</sup>, M. Zhao<sup>1,10</sup>, T. A. ten Brummelaar<sup>8</sup>, J. Sturmann<sup>8</sup> & L. Sturmann<sup>8</sup>

Sunspots are cool areas caused by strong surface magnetic fields that inhibit convection<sup>1,2</sup>. Moreover, strong magnetic fields can alter the average atmospheric structure<sup>3</sup>, degrading our ability to measure stellar masses and ages. Stars that are more active than the Sun have more and stronger dark spots than does the Sun, including on the rotational pole<sup>4</sup>. Doppler imaging, which has so far produced the most detailed images of surface structures on other stars, cannot always distinguish the hemisphere in which the starspots are located, especially in the equatorial region and if the data quality is not optimal<sup>5</sup>. This leads to problems in investigating the north–south distribution of starspot active latitudes (those latitudes with more starspot activity); this distribution is a crucial constraint of dynamo theory. Polar spots, whose existence is inferred from Doppler tomography, could plausibly be observational artefacts<sup>6</sup>. Here we report imaging of the old, magnetically active star  $\zeta$  Andromedae using long-baseline infrared interferometry. In our data, a dark polar spot is seen in each of two observation epochs, whereas lower-latitude spot structures in both hemispheres do not persist between observations, revealing global starspot asymmetries. The north–south symmetry of active latitudes observed on the Sun<sup>7</sup> is absent on  $\zeta$  And, which hosts global spot patterns that cannot be produced by solar-type dynamos<sup>8</sup>.

$\zeta$  And is a nearby active star that is both spatially large and spotted, making it one of a small number of promising targets for imaging with current interferometric capabilities.  $\zeta$  And is a tidally locked close binary (RS CVn) system consisting of a K-type cool giant star and an unseen lower-mass companion star<sup>9</sup>. Tidal interactions have spun-up the cool primary component, causing unusually strong starspots and magnetic activity<sup>4,10</sup>.

We observed  $\zeta$  And during two observing campaigns of eleven nights spanning Universal Time (UT) 9–22 July 2011 and fourteen nights spanning UT 12–30 September 2013 (see Extended Data Table 1) with the Michigan InfraRed Combiner (MIRC)<sup>11</sup> using all six telescopes at Georgia State University's Center for High Angular Resolution Astronomy (CHARA) Array<sup>12</sup> on Mount Wilson, California, USA.

The 2011 and 2013 data sets were separately imaged onto a prolate ellipsoid using the imaging software SURFING (SURFace ImagING), an aperture synthesis imaging technique (J.D.M., manuscript in preparation). This approach replicates the fundamental ideas behind Doppler imaging in that the whole data set is mapped onto the rotating surface at once instead of night-by-night snapshots. Treating each data set as an ensemble also allows SURFING to fit stellar and orbital parameters (see Table 1) along with the surface temperature maps (see Figs 1 and 2).

The surface temperature maps for  $\zeta$  And show peaks of 4,530 K and 4,550 K and minimum values of 3,540 K and 3,660 K in 2011 and 2013, respectively. The  $\sim 900$ -K range of temperatures we see across the surface is slightly larger than the  $\sim 700$ -K range found from recent Doppler imaging work (from the Fe I  $\lambda = 6,430$  Å line). A strong dark polar spot is present in both of our imaging epochs, also consistent with recent Doppler imaging studies<sup>7,13,14</sup>. In contrast to this persistent feature, many other large dark regions change completely between 2011 and 2013 with no apparent overall symmetry or pattern. These features and their locations can only unambiguously be imaged by interferometry, since Doppler and light-curve inversion imaging techniques experience latitude degeneracies (see Methods section for more details). We now discuss the starspot implications on the dynamical large-scale magnetic field of  $\zeta$  And.

The extended network of cool regions stretching across the star suggest that strong magnetic fields can suppress convection on global scales, rather than just local concentrations forming spot structures. The extent to which starspots can cover the surface of a star is at present unknown and of interest in understanding how activity saturates on rapidly rotating, convective stars<sup>15</sup>. The observations in hand lend support to studies that have suggested magnetic activity can be so widespread as to alter the apparent fundamental parameters of a star<sup>16,17</sup>. For example, a larger region of suppressed convection gives a lower observed temperature, leading to inaccurate estimates for stellar mass and age<sup>3</sup>. The changes in global magnetic features will produce long-term photometric variations that are often attributed only to changes in a growing or shrinking polar starspot. We note that a polar starspot for  $\zeta$  And does not affect the flux of the star as much as other large-scale magnetic structures do, owing to the effects

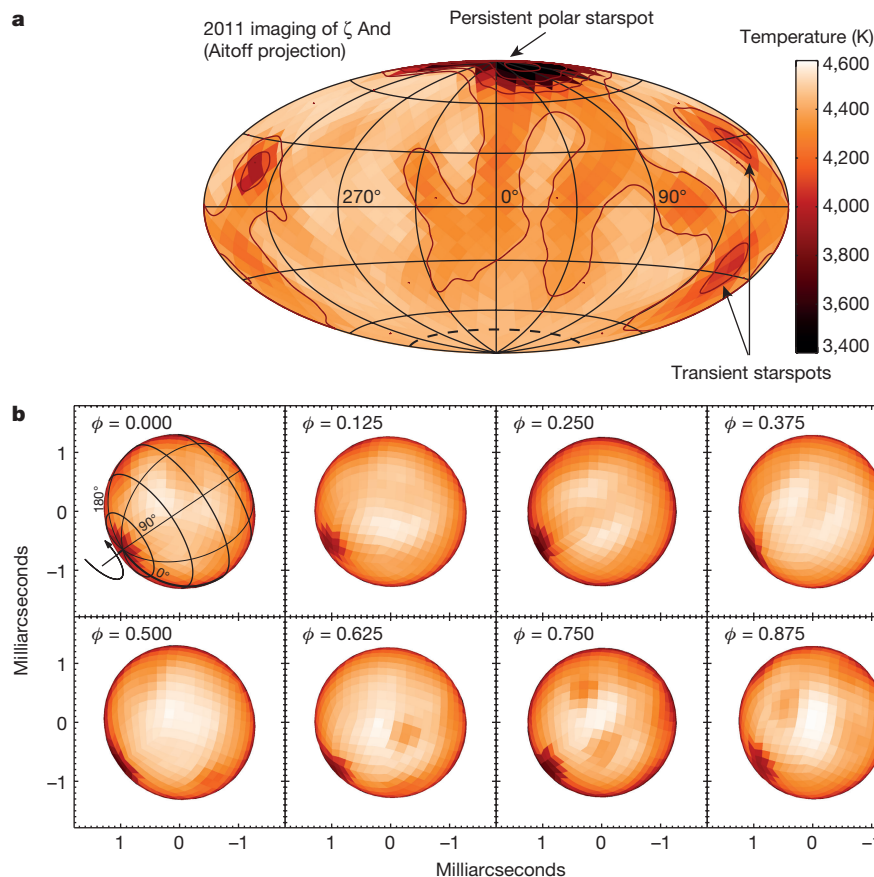
**Table 1 | Parameters of  $\zeta$  And**

Measured parameter	Value
Angular polar diameter, $\theta_{\text{LD}}$ (mas)	$2.502 \pm 0.008$
Polar radius ( $R_{\text{p}}$ )	$15.0 \pm 0.8$
Oblateness (major to polar axis)	$1.060 \pm 0.011$
Inclination, $i$ (°)	$70.0 \pm 2.8$
Pole position angle (°, E of N)	$126.0 \pm 1.9$
<b>Values from the literature</b>	
Distance, $d$ (pc)	$17.98 \pm 0.83$ (ref. 29)
Effective temperature, $T_{\text{eff}}$ (K)	$4600 \pm 100$ (ref. 9)
Luminosity, $\log L/L_{\odot}$	$1.98 \pm 0.04$ (ref. 9)
Primary mass ( $M_{\odot}$ )	$2.6 \pm 0.4$ (ref. 9)
Secondary mass ( $M_{\odot}$ )	$\sim 0.75$ (ref. 9)
Iron metallicity $[\text{Fe}/\text{H}]/[\text{Fe}/\text{H}]_{\odot}$	$-0.30 \pm 0.05$ (ref. 9)

SURFING models assumed a circular orbit (eccentricity  $e = 0$ ) using circular radial velocity conventions with an orbital period  $P_{\text{orb}} = 17.7694260 \pm 0.00004$  days and time of nodal passage  $T_0 = 49992.281 \pm 0.017$  (MJD)<sup>30</sup>. Limb darkening was held fixed with power-law exponent  $\mu = 0.269$ , appropriate for  $\zeta$  And based upon spectral type.

<sup>1</sup>Department of Astronomy, University of Michigan, Ann Arbor, Michigan 48109, USA. <sup>2</sup>Finnish Centre for Astronomy with ESO (FINCA), University of Turku, FI-21500 Piikkiö, Finland. <sup>3</sup>Dark Cosmology Centre, Niels Bohr Institute, University of Copenhagen, Juliane Maries Vej 30, DK-2100 Copenhagen Ø, Denmark. <sup>4</sup>Department of Physics and Astronomy, Georgia State University, Atlanta, Georgia 30303, USA. <sup>5</sup>Department of Physics and Astronomy, Ohio Wesleyan University, Delaware, Ohio 48103, USA. <sup>6</sup>Konkoly Observatory, Research Center for Astronomy and Earth Sciences, Hungarian Academy of Sciences, H-1121 Budapest, Konkoly Thege Miklós út 15-17, Hungary. <sup>7</sup>School of Physics, University of Exeter, Exeter, EX4 4QL, UK. <sup>8</sup>Center for High Angular Resolution Astronomy, Georgia State University, Mount Wilson, California 91023, USA. <sup>9</sup>Harvard-Smithsonian Center for Astrophysics, 60 Garden Street, Cambridge, Massachusetts 02138, USA. <sup>10</sup>Department of Astronomy and Astrophysics, Pennsylvania State University, State College, Pennsylvania 16802, USA.





**Figure 1 | Surface image of  $\zeta$  And from July 2011 with eleven nights of data using SURFING.** **a**, The temperature of  $\zeta$  And is presented in an Aitoff projection. The contours represent every 200 K from 3,400 K to 4,600 K. The dashed line at the bottom pole indicates the latitudes below, which are hidden owing to the inclination. Arrows point to example

transient starspots. **b**, The surface reflects how the star is observed on the sky with  $H$ -band intensities (mean  $H = 1.64$ ). The phase  $\phi = 0.000$  plot shows longitude  $0^\circ$  at the bottom right of the star with  $90^\circ$  across the middle. The phases assume circular orbit radial velocity conventions. The images are oriented with north up and east to the left.

of limb darkening and foreshortening on this highly inclined system ( $i \approx 70.0^\circ$ ).

The interferometric images of  $\zeta$  And provide a clear confirmation of the existence of polar spots. Polar spots have been seen in Doppler images of  $\zeta$  And (refs 9, 13 and 14) and of many other active stars<sup>4</sup>. Polar spots produce spectral line-profile changes only in the line core itself (no Doppler shift), and the spectral signature of a symmetric polar spot is the same at each rotational phase of the star. This signature means that polar spots can very easily be produced as artefacts in the Doppler imaging process; if the depth of the spectral line profile is not correctly modelled, then the image will exhibit a polar spot. Strong chromospheric activity has also been postulated to fill in at least some of the photospheric lines used in Doppler imaging, potentially producing a polar spot<sup>18,19</sup>. These facts made the existence of polar spots a matter of debate in the early days of Doppler imaging<sup>20,21</sup> and this is now independent confirmation of their existence, settling the debate.

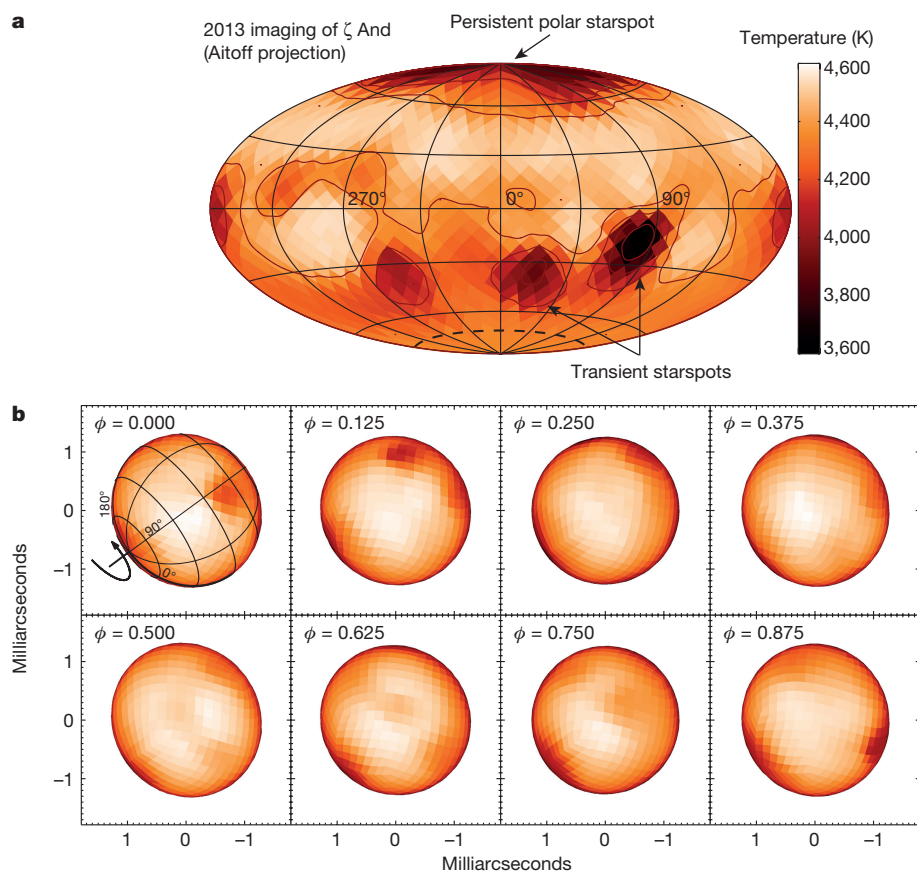
The interferometric images of  $\zeta$  And presented here reveal the exact hemispheres of the spots and show strong asymmetries between the hemispheres, with the 2011 map showing dominant spots on the northern and the 2013 map showing them on the southern hemisphere. On the Sun the spots are typically seen on both hemispheres in certain active mid-latitude regions, with some breaking of the symmetry in the spot numbers on the two hemispheres<sup>7</sup> but not as large an asymmetry as seen on  $\zeta$  And. Such asymmetries require a departure from the solar-type  $\alpha\Omega$ -dynamo to a more complicated dynamo for  $\zeta$  And, such as one with mixed parity modes<sup>22</sup>.

Although our results strictly apply only for giant stars in RS CVn binaries, we note strong parallels between the physical conditions and

magnetic behaviour of these and pre-main-sequence stars. To reach these conditions, the giant primary stars in RS CVn binaries rotate rapidly owing to tidal spin-up and pre-main-sequence stars rotate rapidly owing to contraction and angular momentum transfer caused by accretion of material from a circumstellar disk. These similar physical conditions hint at shared field-generation mechanisms that are observationally indistinguishable<sup>23</sup> and manifest as starspots. In young associations, it has been noted that derived ages are probably strongly affected by global suppression of convection<sup>3</sup>. These commonalities and the known consequences argue that strong stellar magnetism must be accounted for in models for both pre-main-sequence stars and for the stages of evolution in the most active giant stars.

Results from imaging studies using light-curve inversion and Doppler imaging techniques, as well as new interferometric spot studies<sup>24</sup>, all re-enforce the picture that global magnetic structures cover the faces of the most active stars. Our interferometric imaging has found unambiguous signposts of these structures and clearly points to a perspective beyond the typical isolated spots observed on the Sun. The large-scale suppression of convection by these global magnetic fields will have structural effects on the stellar atmosphere, including puffing up the star and decreasing the effective temperature and luminosity, dramatic alterations that must be accounted for by modern stellar structure calculations especially for young, low-mass stars that universally show strong magnetic activity<sup>3,25</sup>.

To understand these structural effects, we must image more targets with as much detail as possible. The procedures used here can provide similar  $H$ -band images for a handful of bright, spatially large, spotted stars (such as  $\sigma$  Geminorum and  $\lambda$  Andromedae). Impending



**Figure 2 | Surface images of  $\zeta$  And from September 2013 with fourteen nights of data using SURFING.** **a** and **b** are presented as in Fig. 1, except that the 200-K contours of the Aitoff projection (**a**) range from 3,600 K to 4,600 K. The polar spot is observed to have evolved between the two sets of

observations. The lower-latitude spots present in the 2011 data set are not present in the 2013 data set, with the new spots located mostly below the equator, emphasizing the spot-latitude asymmetry observed.

advances in visible interferometry will allow for similar resolution on more stars (down to  $\theta \approx 1.1$  mas). For stars that cannot be resolved in detail, combining interferometrically observed photocentre shifts due to rotation of starspots in and out of view with Doppler imaging would resolve the degeneracies inherent in the Doppler images, allowing for more accurate surface maps. By acquiring a number of these maps on several stars or a few observation epochs of the same targets, we could investigate how the changing magnetic field affects our determinations of stellar parameters (including mass and age)<sup>3,26</sup>. In addition, the development of new dynamo models would shed light on the impact of magnetism on stellar evolution<sup>27,28</sup>.

**Online Content** Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

**Received 30 October 2015; accepted 18 February 2016.**

**Published online 4 May 2016.**

- Hale, G. E. On the probable existence of a magnetic field in sun-spots. *Astrophys. J.* **28**, 315–343 (1908).
- Biermann, L. Der gegenwärtige Stand der Theorie konvektiver Sonnenmodelle. *Vierteljahresschrift Astron. Gesells.* **76**, 194–200 (1941).
- Somers, G. & Pinnsonneault, M. H. Older and cooler: the impact of starspots on pre-main-sequence stellar evolution. *Astrophys. J.* **807**, 174 (2015).
- Strassmeier, K. G. Starspots. *Astron. Astrophys. Rev.* **17**, 251–308 (2009).
- Piskunov, N. E., Tuominen, I. & Vilhu, O. Surface imaging of late-type stars. *Astron. Astrophys.* **230**, 363–370 (1990).
- Bruls, J. H. M. J., Solanki, S. K. & Schuessler, M. Doppler imaging: the polar SPOT controversy. *Astron. Astrophys.* **336**, 231–241 (1998).
- Hathaway, D. F. The solar cycle. *Liv. Rev. Sol. Phys.* **12**, 4 (2015).
- Norton, A. A., Charbonneau, P. & Passos, D. Hemispheric coupling: comparing dynamo simulations and observations. *Space Sci. Rev.* **186**, 251–283 (2014).

- Kovári, Zs. *et al.* Doppler imaging of stellar surface structure. XXIII. The ellipsoidal K giant binary  $\zeta$  Andromedae. *Astron. Astrophys.* **463**, 1071–1080 (2007).
- Berdugina, S. Starspots: a key to the stellar dynamo. *Liv. Rev. Sol. Phys.* **2**, 8 (2005).
- Monnier, J. D., Berger, J.-P., Millan-Gabet, R. & ten Brummelaar, T. A. The Michigan Infrared Combiner (MIRC): IR imaging with the CHARA array. *Proc. SPIE* **5491**, 1370–1378 (2004).
- ten Brummelaar, T. A. *et al.* First results from the CHARA Array. II. A description of the instrument. *Astrophys. J.* **628**, 453–465 (2005).
- Korhonen, H. *et al.* Ellipsoidal primary of the RS CVn binary  $\zeta$  Andromedae. Investigation using high-resolution spectroscopy and optical interferometry. *Astron. Astrophys.* **515**, A14 (2010).
- Kovári, Zs. *et al.* Measuring differential rotation of the K-giant  $\zeta$  Andromedae. *Astron. Astrophys.* **539**, A50 (2012).
- Reiners, A., Basri, G. & Browning, M. Evidence for magnetic flux saturation in rapidly rotating M stars. *Astrophys. J.* **692**, 538–545 (2009).
- Spruit, H. C. Effect of spots on a star's radius and luminosity. *Astron. Astrophys.* **108**, 348–355 (1982).
- López-Morales, M. On the correlation between the magnetic activity levels, metallicities, and radii of low-mass stars. *Astrophys. J.* **660**, 732–739 (2007).
- Unruh, Y. C. & Collier Cameron, A. Does chromospheric emission mimic polar starspots in Doppler images? *Mon. Not. R. Astron. Soc.* **290**, L37–L42 (1997).
- Bruls, J. H. M. J., Solanki, S. K. & Schuessler, M. Doppler imaging: the polar SPOT controversy. *Astron. Astrophys.* **336**, 231–241 (1998).
- Strassmeier, K. G. *et al.* Doppler imaging of high-latitude SPOT activity on HD 26337. *Astron. Astrophys.* **247**, 130–147 (1991).
- Piskunov, N. & Wehlau, W. H. The detectability of cool polar caps on late type stars. *Astron. Astrophys.* **289**, 868–870 (1994).
- Bushby, P. J. Strong asymmetry in stellar dynamos. *Mon. Not. R. Astron. Soc.* **338**, 655–664 (2003).
- Bouvier, J. & Bertout, C. Spots on T Tauri stars. *Astron. Astrophys.* **211**, 99–114 (1989).
- Parks, J. R. *et al.* First images of cool starspots on a star other than the Sun: interferometric imaging of  $\lambda$  Andromedae. *Astrophys. J.* (submitted); preprint at <http://arxiv.org/abs/1508.04755> (2016).
- Oláh, K. *et al.* Magnitude-range brightness variations of overactive K giants. *Astron. Astrophys.* **572**, A94 (2014).

26. Stassun, K. G., Kratter, K. M., Scholz, A. & Dupuy, T. J. An empirical correction for activity effects on the temperatures, radii, and estimated masses of low-mass stars and brown dwarfs. *Astrophys. J.* **756**, 47 (2012).
27. MacDonald, J. & Mullan, D. J. Structural effects of magnetic fields in brown dwarfs. *Astrophys. J.* **700**, 387–394 (2009).
28. Chabrier, G., Gallardo, J. & Baraffe, I. Evolution of low-mass star and brown dwarf eclipsing binaries. *Astron. Astrophys.* **472**, L17–L20 (2007).
29. ESA *The Hipparcos and Tycho Catalogues*. ESA SP-1200, <http://www.cosmos.esa.int/web/hipparcos/catalogues> (ESA, 1997).
30. Fekel, F. C., Strassmeier, K. G., Weber, M. & Washuettl, A. Orbital elements and physical parameters of ten chromospherically active binary stars. *Astron. Astrophys. Suppl. Ser.* **137**, 369–383 (1999).

**Acknowledgements** We thank E. Pedretti for his early efforts in imaging  $\zeta$  And. This work is based upon observations obtained with the Georgia State University Center for the High Angular Resolution Astronomy Array at Mount Wilson Observatory. The CHARA Array is supported by the National Science Foundation under grant numbers AST-1211929 and AST-1411654. Institutional support has been provided from the GSU College of Arts and Sciences and the GSU Office of the Vice President for Research and Economic Development. The MIRC instrument at the CHARA Array was funded by the University of Michigan. The 2013 CHARA/MIRC observations were supported by a Rackham Graduate Student Research Grant from the

University of Michigan. This imaging work was supported by the National Science Foundation (NSF) grant AST-1108963. The SURFING software was developed in part with funding from the Observatoire de Paris, Meudon. This research made use of the SIMBAD database, operated at CDS, Strasbourg, France and the Jean-Marie Mariotti Center SearchCal and Aspro2 services co-developed by FIZEAU and LAOG/IPAG. This work was supported by the Hungarian Scientific Research Fund (OTKA K-109276), the 'Lendület-2009' Young Researchers' programme of the Hungarian Academy of Sciences, an STFC Rutherford Fellowship and Grant (ST/J004030/1, ST/K003445/1), and an ERC Starting Grant (grant agreement number 639889).

**Author Contributions** R.M.R., J.D.M., F.B., X.C., S.K. and M.Z. obtained the observations of  $\zeta$  And. R.M.R., J.D.M., F.B., X.C., S.K., G.H.S. and M.Z. obtained the observations of 37 And. R.M.R. performed the data reduction and calibration. R.M.R. and G.T. determined the orbital parameters of 37 And. J.D.M. and R.M.R. created the images of  $\zeta$  And, which were interpreted by R.M.R., J.D.M., H.K., Zs.K., R.O.H. and A.N.A. T.A.t.B., G.H.S., J.S. and L.S. provided observational setup and technical support.

**Author Information** Reprints and permissions information is available at [www.nature.com/reprints](http://www.nature.com/reprints). The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to R.M.R. ([rmroett@umich.edu](mailto:rmroett@umich.edu)).



## METHODS

**Starspot imaging methods.** Large spots covering a substantial fraction of the stellar surface have been indirectly imaged using light-curve inversion and Doppler imaging techniques<sup>31,32</sup>. Light-curve inversion reproduces spotted stellar surfaces based on time-series data and can only reproduce structures observed as rotational modulations—structures such as static polar spots are practically invisible to light-curve inversion imaging. Light-curve inversion typically reveals only weak relative latitude information for the spots, although this can be improved with the combination of concurrent observations in multiple filters<sup>33</sup>. A more detailed surface map, both in latitude and longitude, can be obtained with Doppler imaging, which creates surface temperature maps from tracking small changes in absorption lines as starspots rotate in and out of view<sup>32</sup>. Nevertheless, this method cannot always distinguish the hemisphere in which the structures are located. To confirm important findings from these methods and to understand the global characteristics of activity that can alter stellar radii and effective temperature, a more direct imaging method is required that is immune to these ambiguities.

The nearest magnetically active stars are too small to be resolved by even our largest telescopes. However, long-baseline interferometry has the potential to image sub-milliarcsecond features on the surfaces of nearby stars. To date, interferometric imaging has been used successfully to confirm the oblateness and gravity darkening of rapidly rotating stars<sup>34</sup> and even to image a spotted stellar surface<sup>24</sup>. To improve the resolution and imaging quality for rotating stars that show strong magnetic activity, we present here an ‘imaging-on-a-sphere’ technique that uses interferometric observations from multiple nights to constrain a surface temperature map. This naturally takes advantage of the multiple views we have of starspot structures as they rotate across the disk of the star. Interferometric imaging produces unique images that resolve the degeneracies in latitude of Doppler imaging and provides the first independent confirmation of the existence of polar starspots.

**Interferometric data.** Our interferometric observations were obtained using the Michigan InfraRed Combiner (MIRC) at the Center for High Angular Resolution Astronomy (CHARA) Array. The CHARA Array consists of six 1-m telescopes in a Y-shaped configuration with baselines ranging from 34 m to 331 m (ref. 12). For the  $\zeta$  And data, MIRC<sup>11</sup> combined light from all six CHARA Array telescopes in the *H*-band (eight channels across 1.5–1.8  $\mu$ m for  $\lambda/\Delta\lambda \approx 40$ ), resulting in an angular resolution of  $\lambda/2B \approx 0.5$  milliarcseconds, where  $B$  is the longest baseline used by the interferometer, that is, the distance between the two farthest apart telescopes of the interferometer.

The data were reduced and calibrated with the standard MIRC pipeline<sup>35</sup>. We searched without success for evidence of the faint companion in our interferometric data using our proven grid search method<sup>36</sup>, and could secure only a  $1\sigma$  lower limit of 300:1 on the *H*-band flux ratio between primary and companion.

The data products obtained from reducing the CHARA/MIRC data with the standard pipelines consist of visibilities, closure phases and triple amplitudes. Representative samples of the visibilities and closure phases are presented in Extended Data Figs 1 and 2 for a single night (UT 15 September 2013) of six-telescope CHARA/MIRC observations of  $\zeta$  And. The reduced data are available in OI-FITS format<sup>37</sup> upon request.

**Calibration stars.** The twenty-five nights of interferometric data span 9–22 July 2011 and 12–30 September 2013. For these nights of observation we use four calibration stars (37 Andromedae,  $\gamma$  Pegasi,  $\gamma$  Trianguli and 58 Ophiuchi) interspersed with observations of the target star  $\zeta$  And.  $\gamma$  Peg,  $\gamma$  Tri, and 58 Oph are modelled as spherical, uniform disk stars<sup>38</sup> with their parameters as in Extended Data Table 2.

The calibrator 37 And is a recently discovered binary system<sup>39</sup> with a primary-to-secondary *H*-band flux ratio of  $80 \pm 20$ . Ordinarily, binary stars make poor calibrators, but the 37 And system was observed enough times to determine its orbit precisely and salvage its use for calibrating our primary target  $\zeta$  And. We detect the companion of 37 And in nineteen nights of data (see Extended Data Table 3) using a grid search for the companion. To constrain orbital parameters, we combined the visual orbit with the primary star’s radial velocity curve obtained with archival spectra from the ELODIE high-resolution échelle spectrograph formerly on a 1.93-m telescope at Observatoire de Haute-Provence, France<sup>40</sup>. Extended Data Figs 3 and 4 show the system orbit and radial velocity curve and Extended Data Table 4 contains the system orbital parameters. The orbital parameters are used in the MIRC calibration pipeline to account for the effect of the companion of 37 And.

**SURFING imaging code.** The image reconstruction code SURFING (SURface imagingING) was specially written for this project: to image surfaces of rotating stars. We create a global model of the star, including geometrical parameters (polar radius, oblateness, inclination, pole position angle, limb darkening

coefficient, rotational period, epoch) as well as the surface temperature map. We cover the surface with tiles of equal area using the HEALPix methodology<sup>41</sup>, using 768 tiles to match the spatial resolution of CHARA. Each tile has an area of  $0.025 \text{ mas}^2$ . We represent the shape of the stellar photosphere as a prolate spheroid to approximate a slightly filled Roche potential. The deviation from spherical appears as gravity darkening and accounts for a temperature difference of only  $\sim 60 \text{ K}$ , which is much smaller than the temperature variations of the starspots.

We sampled the large range of geometrical models using an affine invariant ensemble Markov chain Monte Carlo approach<sup>42</sup>, with a nested loop to iteratively optimize the surface temperature map within each walker of the outer loop (this was needed because the 768 free parameters needed to characterize the temperature map would be intractable using a Markov chain approach). Extensive testing was carried out using blind simulated data to optimize the speed of convergence. In addition to minimizing the  $\chi^2$  statistic, we also incorporated priors on each parameter and could experiment with a variety of imaging regularizers, such as total variation or the L2norm of wavelet coefficients.

In addition to testing the code on simulated data, we were also able to check results using soon-to-be-published data on the spotted star  $\lambda$  And (ref. 24), finding comparable results and confirming the inclination and position angle of the pole. We also checked that the imaging-derived orientation of another RS CVn ( $\sigma$  Gem; R.M.R., J.D.M., H.K., R.O.H., F.B., X.C., G. W. Henry, G.H.S., M. Weber & K. G. Strassmeier, manuscript in preparation) matched the orbital plane of the close companion. The results of these tests and additional details about the implementation of our method will be described in a future paper (J.D.M., manuscript in preparation).

**$\zeta$  And parameters from SURFING.** Another test of the robustness of our fitting methodology is to determine the stellar orientation for independent data sets and compare the results. While the magnetic field structures will vary from year to year, the inclination of the pole and its position angle on the sky will not. Extended Data Fig. 5 shows a  $\chi^2$  surface for three separate years: a 2008 pilot set of observations using the MIRC four-telescope system, the 2011 data set, and the 2013 data set. This figure shows two large regions of reduced  $\chi^2$  around inclination  $i = 90^\circ$  and position angle (PA)  $= -60^\circ$  or  $120^\circ$ . These regions reflect the oblateness of the star and the basic orientation on the sky. The region near inclination  $i = 70^\circ$ , PA  $= 120^\circ$  has the lowest  $\chi^2$  in all years, and especially in 2013. This tells us that the spots move from the southwest towards the northeast and not the other way around, and the consistent picture from year to year both shows the efficacy of the code and gives confidence that we are measuring the true astrophysical signal and not over-fitting noise or systematic errors.

The results from these grid studies have been used to robustly estimate the geometrical parameters for  $\zeta$  And, and these parameters are found in Table 1. The error bars associated with the parameters were determined by combining the results of data sets from 2011 and 2013, with the additional data from 2008. The 2008 data were obtained with only four CHARA telescopes, so they are not of high enough quality for imaging.

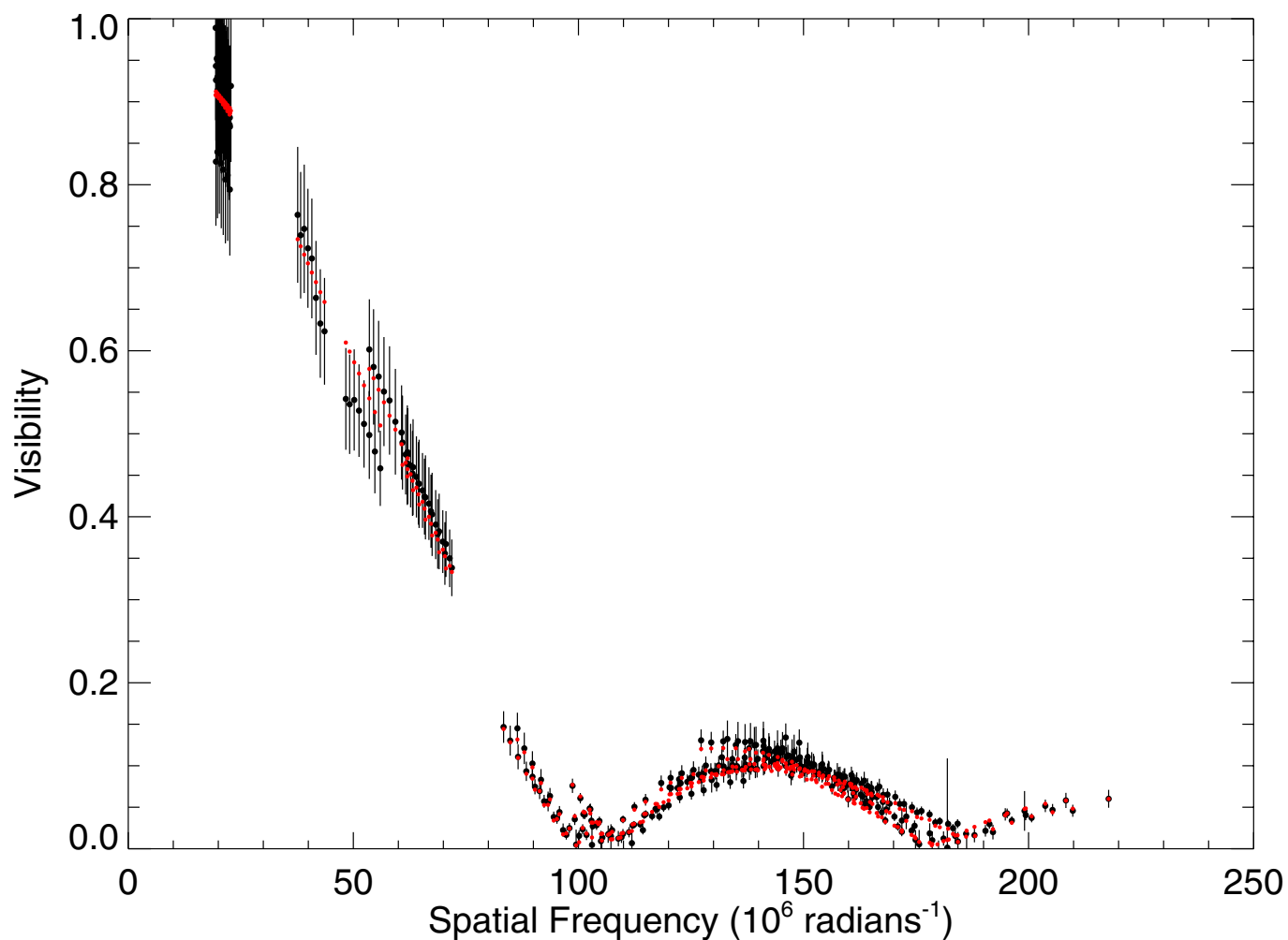
To convert *H*-band intensities from the reconstructed images into photospheric temperatures, we used Kurucz atmospheric models<sup>43</sup> for iron metallicity  $[\text{Fe}/\text{H}] = -0.25$  and appropriate surface gravity (logg). Note that the overall temperature scale in our maps is uncertain (overall multiplicative scaling) owing to lack of coeval photometry at *H* band; here we adopted a mean *H*-band magnitude of  $H = 1.64$ , based on archival infrared photometry.

**$\zeta$  And imaging tests.** The previous section laid out our robust method for determining the geometrical parameters of  $\zeta$  And by using three independent observing data from different years. The next issue is to determine the reliability of the surface temperature maps. We split the extensive CHARA/MIRC data from 2013 into two sets of seven nights each, alternating chronologically which night went to which set. Since spots take many days to cross the face of the star, each data set should be viewing the same spots even though the timing was different. We present a comparison of the SURFING results in Extended Data Fig. 6. Both partial data sets reproduce the main features observed by the full data set shown in the main text. This proves that the features seen in the maps are real and are not the result of an over-fitting to poor-quality data or peculiarities of the nightly baseline coverage.

The unprecedented phase coverage in the 2011 and (especially) the 2013 observing runs have allowed for textbook imaging fidelity tests for  $\zeta$  And. We show that the large-scale dark regions that cover  $\zeta$  And are highly robust, probably deriving from magnetically suppressed convection.

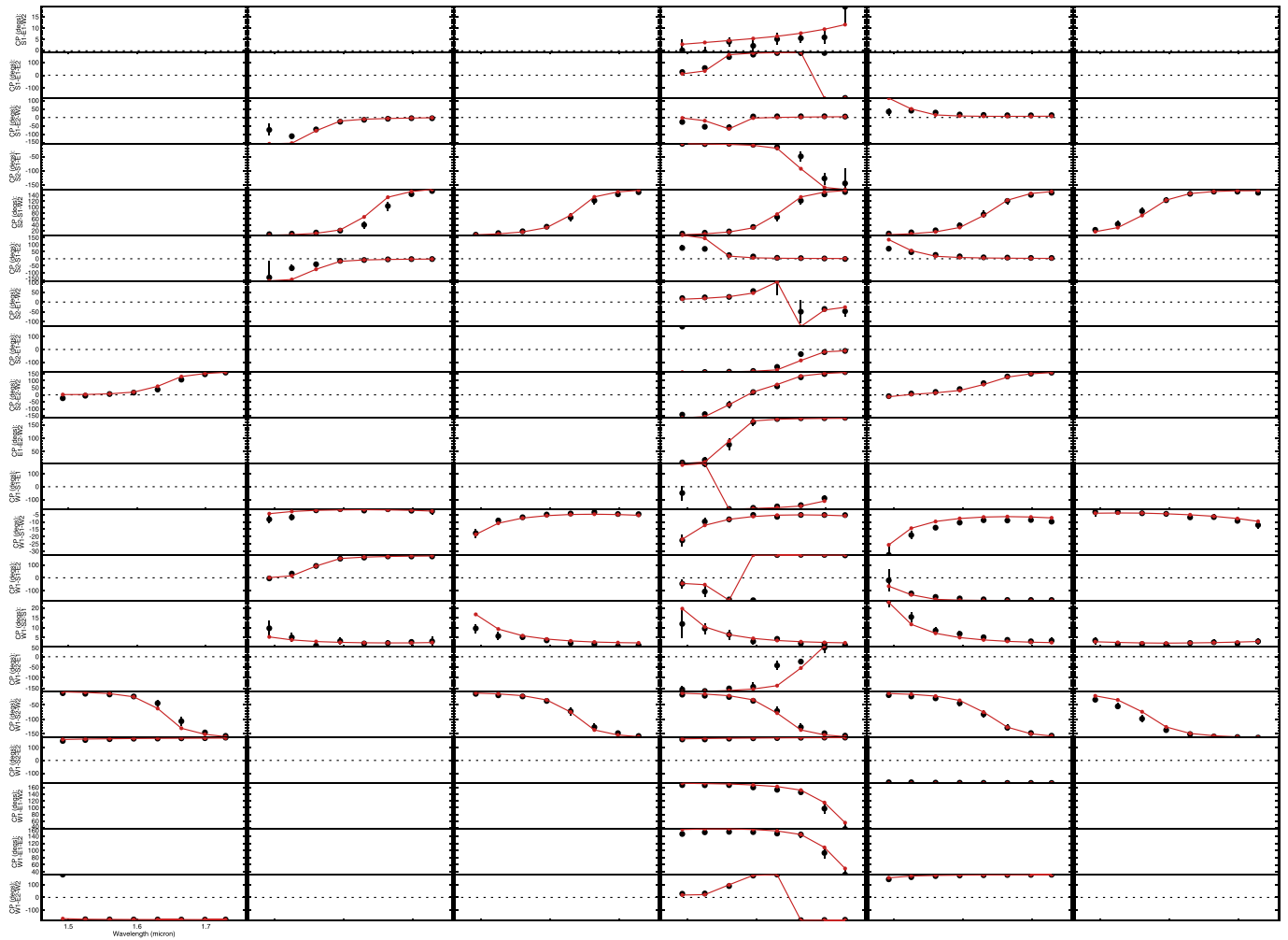
**Code availability.** At present, we have opted not to make the SURFING code available because of a publication in preparation, which will detail the use and the applicability of the resource.

31. Kjurkchieva, D. P. New analysis of II Peg 1977 light curve. *Astrophys. Space Sci.* **155**, 125–129 (1989).
32. Vogt, S. S. & Penrod, G. D. Doppler imaging of spotted stars—application to the RS Canum Venaticorum star HR 1099. *Publ. Astron. Soc. Pacif.* **95**, 565–576 (1983).
33. Harmon, R. O. & Crews, L. J. Imaging stellar surfaces via matrix light-curve inversion. *Astron. J.* **120**, 3274–3294 (2000).
34. Monnier, J. D. *et al.* Imaging the surface of Altair. *Science* **317**, 342–345 (2007).
35. Monnier, J. D. *et al.* Resolving Vega and the inclination controversy with CHARA/MIRC. *Astrophys. J.* **761**, L3 (2012).
36. Roettenbacher, R. M. *et al.* Detecting the companions and ellipsoidal variations of RS CVn primaries. I.  $\sigma$  Geminorum. *Astrophys. J.* **807**, 23 (2015).
37. Pauls, T. A., Young, J. S., Cotton, W. D. & Monnier, J. D. A data exchange standard for optical (visible/IR) interferometry. *Publ. Astron. Soc. Pacif.* **117**, 1255–1262 (2005).
38. Bonneau, D. *et al.* SearchCal: a virtual observatory tool for searching calibrators in optical long baseline interferometry. I. The bright object case. *Astron. Astrophys.* **456**, 789 (2006).
39. Baron, F. *et al.* CHARA/MIRC observations of two M supergiants in Perseus OB1: temperature, Bayesian modeling, and compressed sensing imaging. *Astrophys. J.* **785**, 46 (2014).
40. Moutaka, J., Ilovaisky, S. A., Prugniel, P. & Soubiran, C. The ELODIE archive. *Publ. Astron. Soc. Pacif.* **116**, 693–698 (2004).
41. Górski, K. M., Hivon, E., Banday, A. J., Wandelt, B. D., Hansen, F. K., Reinecke, M. & Bartelmann, M. HEALPix: A framework for high-resolution discretization and fast analysis of data distributed on the sphere. *Astrophys. J.* **622**, 759–771 (2005).
42. Foreman-Mackey, D., Hogg, D. W., Lang, D. & Goodman, J. emcee: the MCMC hammer. *Pub. Astron. Soc. Pacif.* **125**, 306–312 (2013).
43. Kurucz, R. L. Model atmospheres for G, F, A, B, and O stars. *Astrophys. J. Suppl. Ser.* **40**, 1–340 (1979).

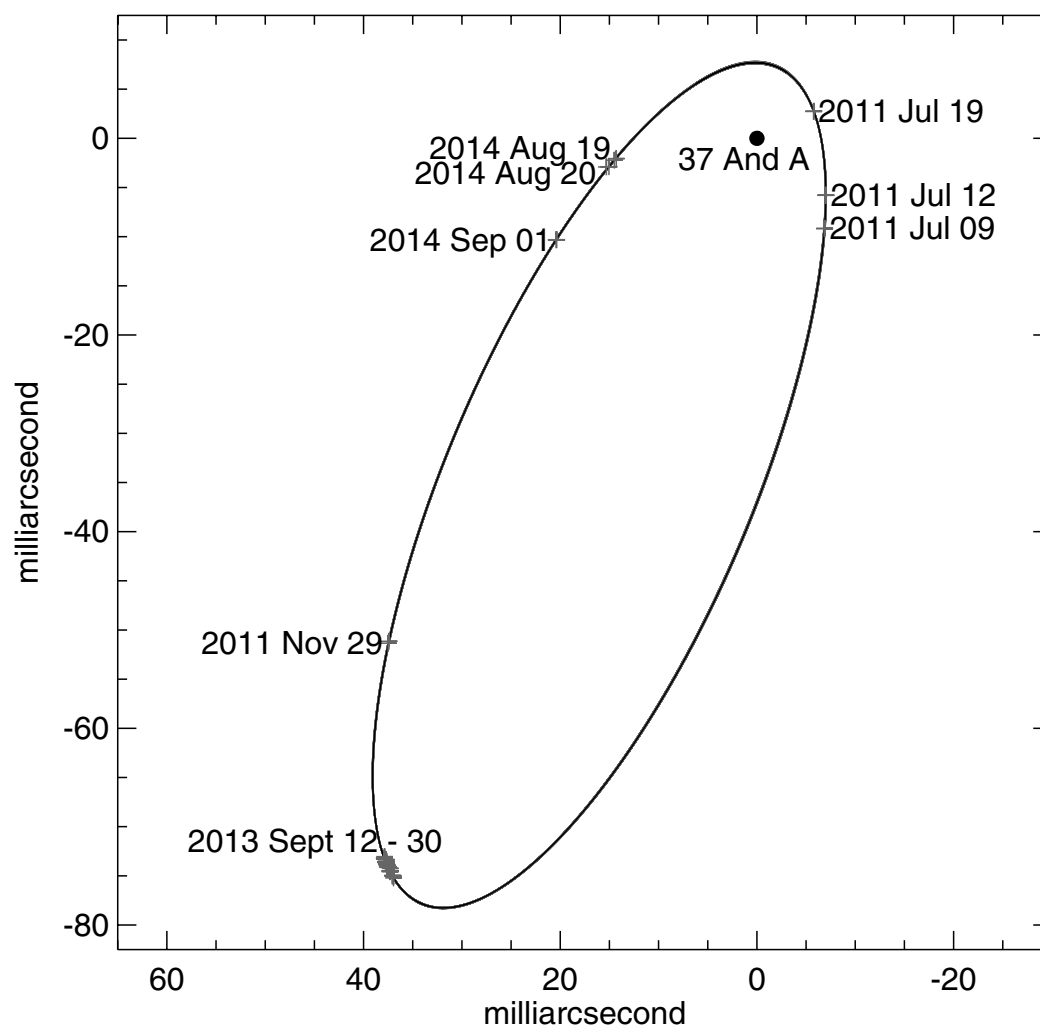


**Extended Data Figure 1 |** Visibility curve of UT 15 September 2013 observations of  $\zeta$  And with CHARA/MIRC. The observed visibilities (unitless) are plotted in black with  $1\sigma$  error bars and the SURFING model visibilities are overlaid in red.

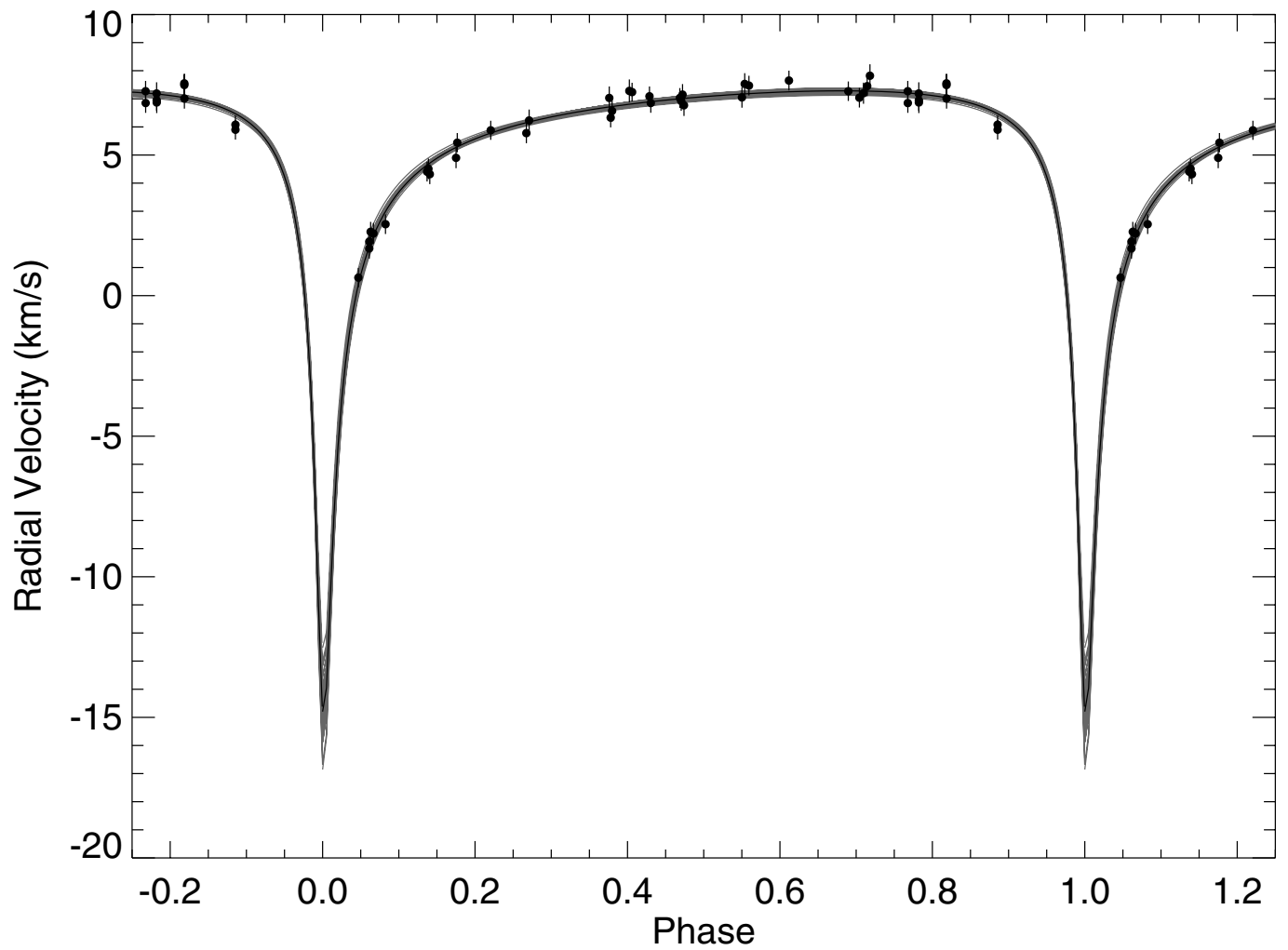




**Extended Data Figure 2 | Closure phases of  $\zeta$  And with CHARA/MIRC.** Each block represents a temporal block of closure phase (CP) observations with data plotted in black (with  $1\sigma$  errors) and SURFING model in red. Each row represents the CP from a unique set of three telescopes, labelled in the standard CHARA format.

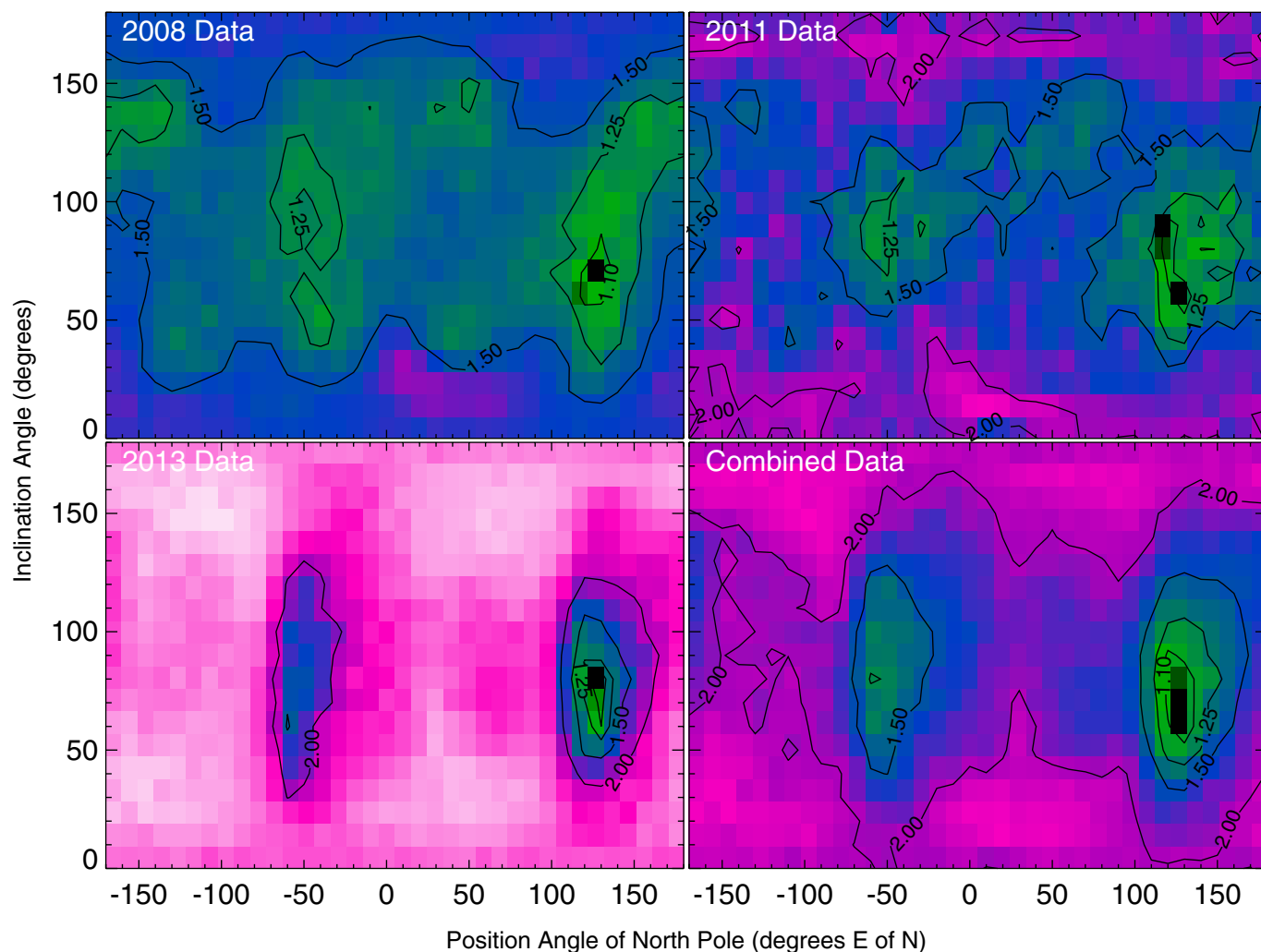


**Extended Data Figure 3 | Orbit of 37 And.** The grey plus signs represent measurements of the companion ( $1\sigma$  errors on detections are smaller than the symbols). The observed resolved disk of 37 And is plotted as the black dot at the origin. The thin solid black line is the best-fit orbit from combining the interferometric detections and the ELODIE radial velocities. Note that the axis units are milliarcseconds (mas) with north up and east to the left.

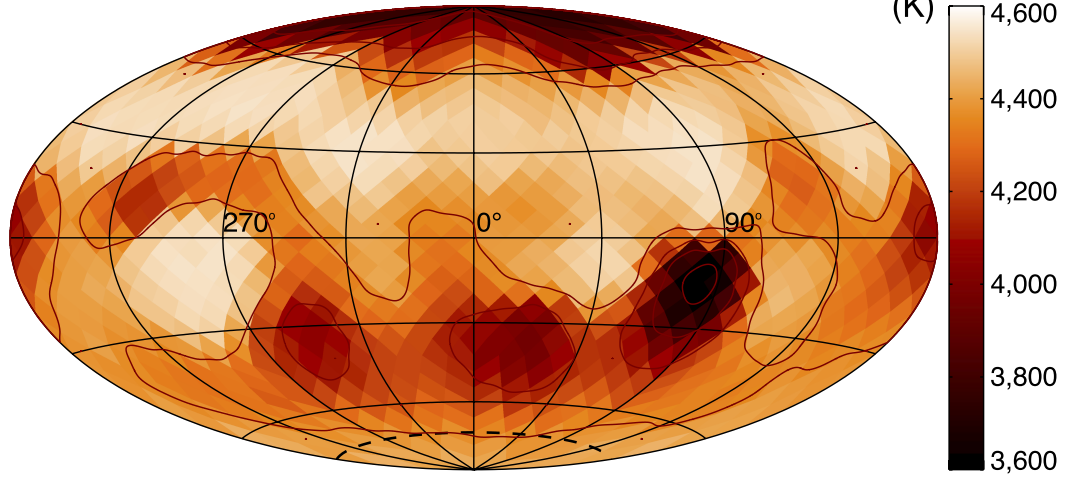
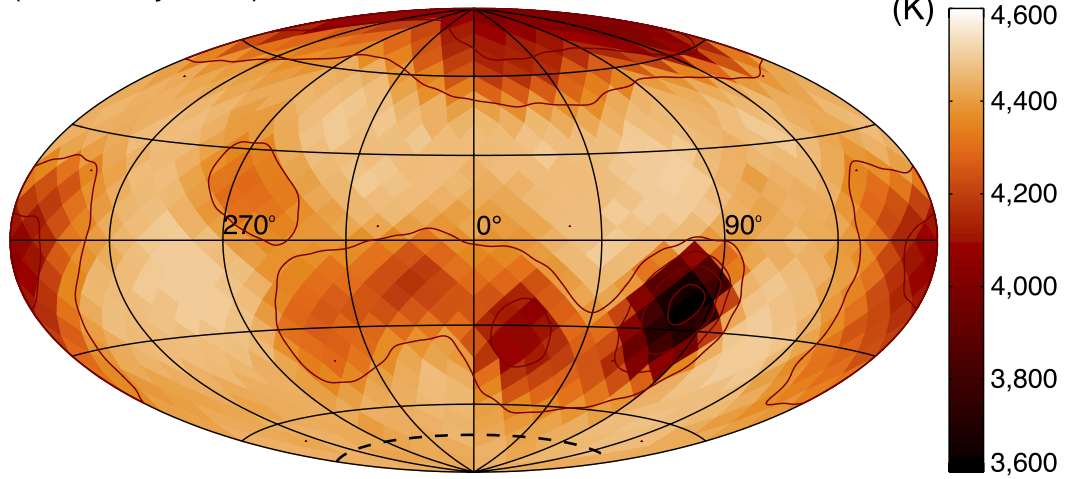


**Extended Data Figure 4 | Radial velocity curve of the primary star of 37 And.** The data points are based upon archival ELODIE spectra (with  $1\sigma$  error bars). The orbital solution used the velocity measurements and the interferometric measurements simultaneously. The solid line is the best-fit orbit and the grey lines are fifty Monte Carlo realizations of the orbit.





**Extended Data Figure 5 | Reduced, normalized  $\chi^2$  surface plot of the SURFING model fitted for position angle and inclination of  $\zeta$  And.** The peak is found at  $PA = 126.0^\circ \pm 1.9$  and  $i = 70.0^\circ \pm 2.8$ . The epoch of each data set is located in the upper left of each panel. The different colours represent varying normalized  $\chi^2$  values with the black contours labelled with specific values.

**a** 2013 Imaging of  $\zeta$  And (Subset A)  
(Aitoff Projection)**b** 2013 Imaging of  $\zeta$  And (Subset B)  
(Aitoff Projection)

**Extended Data Figure 6 | Comparison of the 2013 CHARA/MIRC data set divided into two sets of seven nights of data.** These projections are plotted in the same way as the Aitoff projection in Fig. 2.

Extended Data Table 1 | Observations and calibrators of  $\zeta$  And

UT Date	Modified Julian Date (MJD)	Calibrators Used
2011 Jul 9	55751.536	37 And
2011 Jul 10	55752.531	$\gamma$ Peg
2011 Jul 11	55753.480	37 And
2011 Jul 12	55754.469	$\gamma$ Peg
2011 Jul 14	55756.478	$\gamma$ Peg
2011 Jul 16	55758.505	58 Oph
2011 Jul 17	55759.481	$\gamma$ Peg
2011 Jul 19	55761.475	37 And, $\gamma$ Peg
2011 Jul 20	55762.517	$\gamma$ Peg
2011 Jul 21	55763.478	$\gamma$ Peg, $\gamma$ Tri
2011 Jul 22	55764.480	$\gamma$ Peg, $\gamma$ Tri
2013 Sep 12	56547.449	37 And, $\gamma$ Tri
2013 Sep 13	56548.426	37 And, $\gamma$ Tri
2013 Sep 15	56550.392	37 And, $\gamma$ Peg, $\gamma$ Tri
2013 Sep 16	56551.365	37 And, $\gamma$ Peg, $\gamma$ Tri
2013 Sep 17	56552.345	37 And, $\gamma$ Tri
2013 Sep 18	56553.359	37 And, $\gamma$ Peg, $\gamma$ Tri
2013 Sep 19	56554.407	37 And, $\gamma$ Peg, $\gamma$ Tri
2013 Sep 20	56555.365	37 And, $\gamma$ Peg, $\gamma$ Tri
2013 Sep 21	56556.403	37 And, $\gamma$ Tri
2013 Sep 23	56558.403	37 And, $\gamma$ Tri
2013 Sep 24	56559.343	37 And, $\gamma$ Peg, $\gamma$ Tri
2013 Sep 28	56563.367	37 And, $\gamma$ Tri
2013 Sep 29	56564.357	37 And, $\gamma$ Tri
2013 Sep 30	56565.334	37 And, $\gamma$ Tri



Extended Data Table 2 | Uniform disk sizes (*H*-band) of calibrators

Star Name (HD number)	$\theta_{\text{UD}}$ (mas)
$\gamma$ Peg (HD 886)	$0.41 \pm 0.03$
$\gamma$ Tri (HD 14055)	$0.51 \pm 0.03$
58 Oph (HD 160915)	$0.68 \pm 0.05$

The uniform disk diameters  $\theta_{\text{UD}}$  were obtained with SearchCal<sup>38</sup>.

Extended Data Table 3 | Binary separation and position angle measurements of 37 And

UT Date	Modified Julian Date (MJD)	Separation (mas)	Position Angle (°, E of N)	Error Ellipse Major Axis (mas)	Error Ellipse Minor Axis (mas)	Error Ellipse Position Angle (°)
2011 Jul 9	55751.492	11.46	217.0	0.13	0.11	300
2011 Jul 12	55754.499	9.09	230.6	0.10	0.08	270
2011 Jul 19	55761.465	6.38	295.4	0.06	0.05	330
2011 Nov 29	55894.159	63.57	143.9	1.82	0.16	340
2013 Sep 12	56547.470	82.28	152.6	0.58	0.15	330
2013 Sep 13	56548.399	82.51	152.7	0.02	0.10	30
2013 Sep 15	56550.340	82.47	152.8	0.81	0.26	350
2013 Sep 16	56551.344	82.73	152.9	1.23	0.27	340
2013 Sep 17	56552.329	82.63	152.9	0.48	0.17	330
2013 Sep 18	56553.250	82.88	153.0	0.95	0.21	340
2013 Sep 19	56554.292	83.06	153.1	0.72	0.27	350
2013 Sep 20	56555.301	83.02	153.2	0.79	0.14	330
2013 Sep 23	56558.416	83.08	153.4	1.01	0.22	340
2013 Sep 24	56559.239	83.29	153.5	0.58	0.23	340
2013 Sep 28	56563.303	83.62	153.7	1.08	0.21	340
2013 Sep 30	56565.276	83.65	153.8	0.83	0.20	340
2014 Aug 19	56888.422	14.45	98.2	0.32	0.18	280
2014 Aug 20	56889.446	15.58	100.8	0.24	0.17	10
2014 Sep 1	56901.412	22.82	116.9	0.11	0.07	300

Extended Data Table 4 | Orbital parameters of 37 And

Parameter	Value
Semimajor axis, $a$ (mas)	$46.66 \pm 0.06$
Eccentricity, $e$	$0.8405 \pm 0.0009$
Inclination, $i$ ( $^{\circ}$ )	$52.5 \pm 0.3$
Argument of periastron, $\omega$ ( $^{\circ}$ )	$168.9 \pm 0.3$
Ascending node, $\Omega$ ( $^{\circ}$ )	$-17.6 \pm 0.2$
Orbital period, $P_{\text{orb}}$ (days)	$550.7 \pm 0.2$
Time of periastron passage, $T_0$ (MJD)	$55765.45 \pm 0.04$
Velocity semi-amplitude, $K_A$ (km/s)	$11.1 \pm 0.5$
System velocity, $\gamma$ (km/s)	$5.33 \pm 0.07$

The orbital parameters were obtained by combining the ELODIE radial velocity curve with the CHARA/MIRC detections.



# Temperate Earth-sized planets transiting a nearby ultracool dwarf star

Michaël Gillon<sup>1</sup>, Emmanuël Jehin<sup>1</sup>, Susan M. Lederer<sup>2</sup>, Laetitia Delrez<sup>1</sup>, Julien de Wit<sup>3</sup>, Artem Burdanov<sup>1</sup>, Valérie Van Grootel<sup>1</sup>, Adam J. Burgasser<sup>4</sup>, Amaury H. M. J. Triaud<sup>5</sup>, Cyrielle Opitom<sup>1</sup>, Brice-Olivier Demory<sup>6</sup>, Devendra K. Sahu<sup>7</sup>, Daniella Bardalez Gagliuffi<sup>4</sup>, Pierre Magain<sup>1</sup> & Didier Queloz<sup>6</sup>

Star-like objects with effective temperatures of less than 2,700 kelvin are referred to as ‘ultracool dwarfs’<sup>1</sup>. This heterogeneous group includes stars of extremely low mass as well as brown dwarfs (substellar objects not massive enough to sustain hydrogen fusion), and represents about 15 per cent of the population of astronomical objects near the Sun<sup>2</sup>. Core-accretion theory predicts that, given the small masses of these ultracool dwarfs, and the small sizes of their protoplanetary disks<sup>3,4</sup>, there should be a large but hitherto undetected population of terrestrial planets orbiting them<sup>5</sup>—ranging from metal-rich Mercury-sized planets<sup>6</sup> to more hospitable volatile-rich Earth-sized planets<sup>7</sup>. Here we report observations of three short-period Earth-sized planets transiting an ultracool dwarf star only 12 parsecs away. The inner two planets receive four times and two times the irradiation of Earth, respectively, placing them close to the inner edge of the habitable zone of the star<sup>8</sup>. Our data suggest that 11 orbits remain possible for the third planet, the most likely resulting in irradiation significantly less than that received by Earth. The infrared brightness of the host star, combined with its Jupiter-like size, offers the possibility of thoroughly characterizing the components of this nearby planetary system.

TRAPPIST<sup>9,10</sup> (the TRansiting Planets and Planesimals Small Telescope) monitored the brightness of the star TRAPPIST-1 (2MASS J23062928 – 0502285) in the very-near infrared (roughly 0.9  $\mu\text{m}$ ) at high cadence (approximately 1.2 minutes) for 245 hours over 62 nights from 17 September to 28 December 2015. The resulting light curves show 11 clear transit-like signatures with amplitudes close to 1% (Extended Data Figs 1, 2). Photometric follow-up observations were carried out in the visible range with the Himalayan Chandra 2-metre Telescope (HCT) in India, and in the infrared range with the 8-metre Very Large Telescope (VLT) in Chile and the 3.8-metre UK Infrared Telescope (UKIRT) in Hawaii. These extensive data show that nine of the detected signatures can be attributed to two planets, TRAPPIST-1b and TRAPPIST-1c, transiting the star every 1.51 days and 2.42 days, respectively (Fig. 1a, b). We attribute the two additional transit signals to a third transiting planet, TRAPPIST-1d, for which 11 orbital periods—from 4.5 days to 72.8 days—are possible on the basis of non-continuous observations (Table 1). We cannot discard the possibility that the two transits attributed to planet TRAPPIST-1d originate instead from two different planets, but the consistency of their main parameters (duration, depth and impact parameter) as derived from their individual analyses does not favour this scenario.

TRAPPIST-1 is a well characterized, isolated M8.0  $\pm$  0.5-type dwarf star<sup>11</sup> at a distance of 12.0  $\pm$  0.4 parsecs from Earth as measured by its trigonometric parallax<sup>12</sup>, with an age constrained to be more than 500 million years (Myr), and with a luminosity, mass and radius of 0.05%, 8% and 11.5% those of the Sun<sup>13</sup>, respectively. We determined its metallicity to be solar through the analysis of newly acquired infrared

spectra. The small size of the host star—only slightly larger than Jupiter—translates into Earth-like radii for the three discovered planets, as deduced from their transit depths. Table 1 presents the physical properties of the system, as derived through a global Bayesian analysis of the transit photometry (Fig. 1), including the *a priori* knowledge of its stellar properties, with an adaptive Markov chain Monte Carlo (MCMC) code<sup>14</sup>.

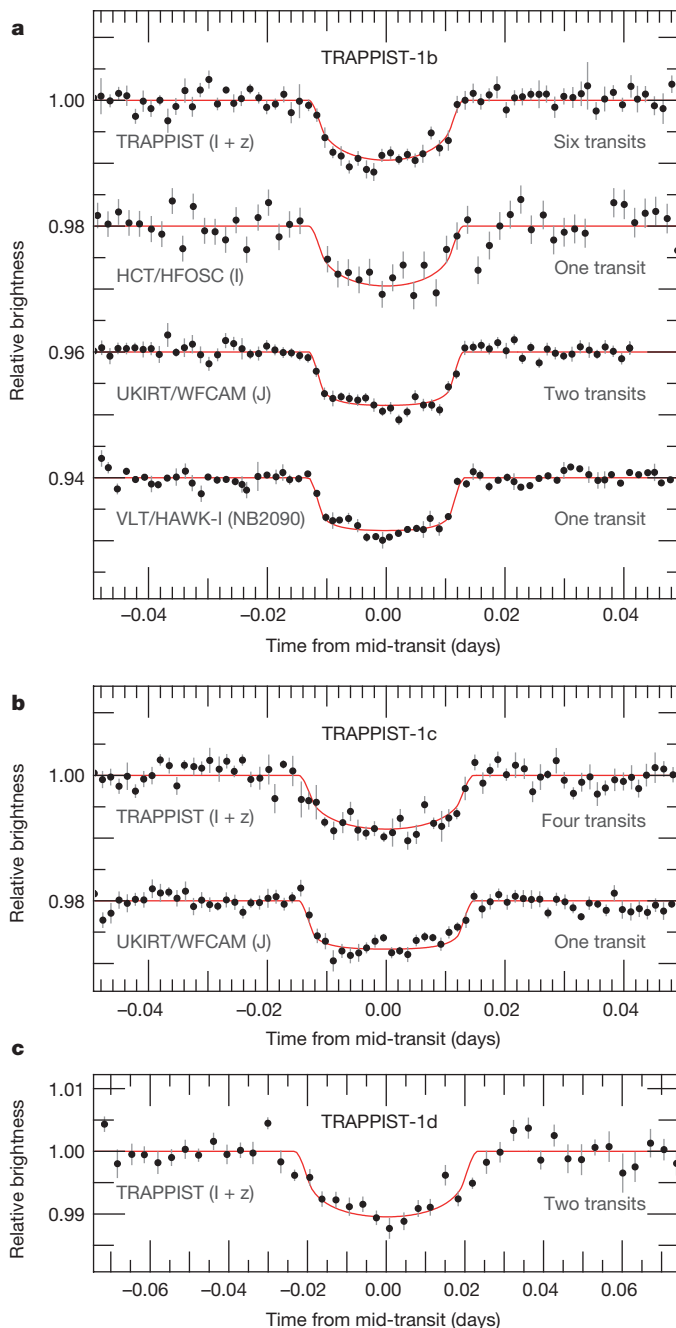
We can discard a non-planetary origin of the transit-like signals owing to several factors. The first factor is the high proper motion of the star (greater than 1'' per year), which allowed confirmation (through archival images) that no background source of significant brightness was located behind it in 2015. The second factor is that the star has no physical companion of stellar-like nature (star or brown dwarf), as demonstrated by high-resolution images, radial velocities and near-infrared spectroscopy. Together, these factors show that the signals do not originate from eclipses of larger bodies in front of a background or a physically associated stellar-like object blended with the ultracool target star. These factors also establish that the light from the target is not diluted by an unresolved additional stellar-like object, confirming that the measured transit depths reveal planetary radii of terrestrial sizes. Other factors include the significant age of the star<sup>13</sup>, its moderate activity<sup>15</sup> and rotation period ( $P_{\text{rot}} = 1.40 \pm 0.05$  days, as measured from our photometry), and its low level of photometric variability<sup>16</sup> (confirmed by our data), all of which are inconsistent with exotic scenarios based on ultrafast rotation of photospheric structures, or on occultations by circumstellar material of non-planetary origin (for example, disk patches or comets)<sup>17</sup>.

Further confirmation of the planetary origin of the transits comes from, first, the periodicity of the transits of TRAPPIST-1b and TRAPPIST-1c, and the achromaticity of the transits of TRAPPIST-1b as observed from 0.85  $\mu\text{m}$  (HCT) to 2.09  $\mu\text{m}$  (VLT) (Fig. 1a); and second, the agreement between the stellar density measured from the transit light curves,  $49.3^{+4.1}_{-8.3} \rho_{\odot}$ , with the density inferred from the stellar properties,  $(55.3 \pm 12.1) \rho_{\odot}$  (where  $\rho_{\odot}$  is the density of the Sun).

The masses of the planets, and thus their compositions, remain unconstrained by these observations. The results of planetary thermal evolution models—and the intense extreme-ultraviolet (1–1,000 Å) emission of low-mass stars<sup>18</sup> during their early lives—make it unlikely that such small planets would have thick envelopes of hydrogen and/or helium gases<sup>19</sup>. Statistical analyses of sub-Neptune-sized planets detected by the Kepler spacecraft indicate that most Earth-sized planets in close orbit around solar-type stars are rocky<sup>20,21</sup>. Nonetheless, the paucity of material in the inner region of the protoplanetary disk of an ultracool dwarf would seem to challenge the *in situ* formation of rocky planets the size of Earth<sup>6</sup>, favouring instead compositions dominated by ice-rich material originating from beyond the ice line<sup>7</sup>. Confirming this hypothesis will require precise mass measurements

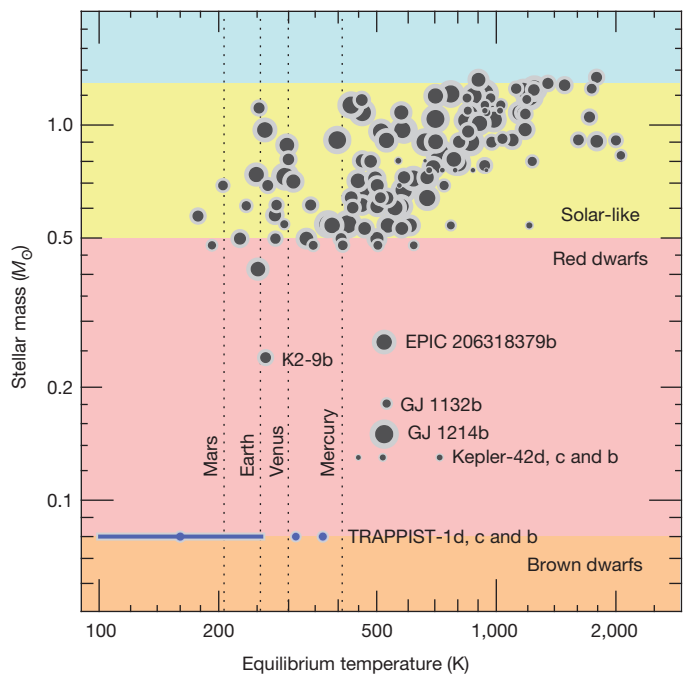
<sup>1</sup>Institut d'Astrophysique et de Géophysique, Université de Liège, Allée du 6 Août 19C, 4000 Liège, Belgium. <sup>2</sup>NASA Johnson Space Center, 2101 NASA Parkway, Houston, Texas, 77058, USA.

<sup>3</sup>Department of Earth, Atmospheric and Planetary Sciences, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, Massachusetts 02139, USA. <sup>4</sup>Center for Astrophysics and Space Science, University of California San Diego, La Jolla, California 92093, USA. <sup>5</sup>Institute of Astronomy, Madingley Road, Cambridge CB3 0HA, UK. <sup>6</sup>Astrophysics Group, Cavendish Laboratory, 19 J J Thomson Avenue, Cambridge, CB3 0HE, UK. <sup>7</sup>Indian Institute of Astrophysics, Koramangala, Bangalore 560 034, India.



**Figure 1 | Transit photometry of the TRAPPIST-1 planets.** Each light curve is phased to the time of inferior conjunction (mid-transit time) of the object. The light curves are binned in two-minute intervals for planet TRAPPIST-1b (a), and in five-minute intervals for planets TRAPPIST-1c (b) and TRAPPIST-1d (c). The best-fit transit models, as derived from a global analysis of the data, are overplotted (red lines). The light curves are shifted along the y axis for the sake of clarity. For the HCT/Hanle faint object spectrograph camera (HFOSC) light curve, the data are unbinned and the error bars are the formal measurement errors. For the other light curves, the error bars are the standard errors of the mean of the measurements in the bin. WFCAM, wide-field infrared camera on the UKIRT; HAWK-I, high acuity wide field K-band imager on the VLT.

so as to break the degeneracy between the relative amounts of iron, silicates and ice<sup>22</sup>. This should be made possible by next-generation, high-precision infrared velocimeters able to measure the low-amplitude Doppler signatures (of one-half to a few metres per second) of the planets. Alternatively, the planets' masses could be constrained by measuring the transit timing variations (TTVs) caused by their



**Figure 2 | Masses of host stars and equilibrium temperatures of known sub-Neptune-sized exoplanets.** The size of the symbols scales linearly with the radius of the planet. The background is colour-coded according to stellar mass (in units of the Sun's mass). The TRAPPIST-1 planets are at the boundary between planets associated with hydrogen-burning stars and planets associated with brown dwarfs. Equilibrium temperatures are estimated neglecting atmospheric effects and assuming an Earth-like albedo of 0.3. The positions of the Solar System terrestrial planets are shown for reference. The range of possible equilibrium temperatures of TRAPPIST-1d is represented by a solid bar; the dot indicates the most likely temperature. Only the exoplanets with a measured radius equal to or smaller than that of GJ 1214b are included.

mutual gravitational interactions<sup>23</sup>, or by transit transmission spectroscopy<sup>24</sup>.

Given their short orbital distances, it is likely that the planets are tidally locked—that is, that their rotations have been synchronized with their orbits by tidal interactions with the host star<sup>25</sup>. Planets TRAPPIST-1b and TRAPPIST-1c are not in the host star's habitable zone<sup>8</sup> (within 0.024 to 0.049 astronomical units (AU) of the star, as defined by one-dimensional models that are not adequate for modelling the highly asymmetric climate of tidally locked planets<sup>26</sup>). However, they have low enough equilibrium temperatures that they might have habitable regions—in particular, at the western terminators of their day sides<sup>27</sup> (Fig. 2 and Table 1). The main concern regarding localized habitability on tidally locked planets relates to the trapping of atmosphere and/or water on their night sides<sup>26</sup>. Nevertheless, the relatively large equilibrium temperatures of TRAPPIST-1b and TRAPPIST-1c would probably prevent such trapping<sup>27</sup>. In contrast, TRAPPIST-1d orbits within or beyond the habitable zone of the star, its most likely periods corresponding to semi-major axes of between 0.033 AU and 0.093 AU. We estimate tidal circularization timescales for TRAPPIST-1d (unlike for the two inner planets) to be more than 1 billion years (see 'Dynamics of the system' in Methods). Tidal heating due to a non-zero orbital eccentricity could thus have a significant influence on the global energy budget and potential habitability of this planet<sup>28</sup>.

The planets' atmospheric properties, and thus their habitability, will depend on several unknown factors. These include the planets' compositions; their formation and dynamical history (their migration and tides); the past evolution and present level of the extreme-ultraviolet stellar flux<sup>29</sup> (probably strong enough in the past, and perhaps even now, to significantly alter the planets' atmospheric compositions<sup>30</sup>);

**Table 1 | Properties of the TRAPPIST-1 planetary system**

Parameter	Value		
<b>Star</b>	<b>TRAPPIST-1 = 2MASS J23062928 – 0502285</b>		
Magnitudes	$V = 18.80 \pm 0.08$ , $R = 16.47 \pm 0.07$ , $I = 14.0 \pm 0.1$ , $J = 11.35 \pm 0.02$ , $K = 10.30 \pm 0.02$		
Distance, $d_*$	$12.1 \pm 0.4$ parsecs (ref. 12)		
Luminosity, $L_*$	$(0.000525 \pm 0.000036)L_\odot$ (ref. 13)		
Mass, $M_*$	$(0.080 \pm 0.009)M_\odot$		
Radius, $R_*$	$(0.117 \pm 0.004)R_\odot$		
Density, $\rho_*$	$50.3^{+5.7}_{-3.3}\rho_\odot$		
Effective temperature, $T_{\text{eff}}$	$2,550 \pm 55$ K		
Metallicity, [Fe/H]	$+0.04 \pm 0.08$ (from near-infrared spectroscopy)		
Rotation period, $P_{\text{rot}}$	$1.40 \pm 0.05$ days (from TRAPPIST photometry)		
Age, $\tau_*$	$>500$ Myr (ref. 13)		
<b>Planets</b>	<b>TRAPPIST-1b</b>	<b>TRAPPIST-1c</b>	<b>TRAPPIST-1d</b>
Orbital period, $P$	$1.510848 \pm 0.000019$ days	$2.421848 \pm 0.000028$ days	$4.551, 5.200, 8.090, 9.101, 10.401, 12.135, 14.561, \mathbf{18.202}, 24.270, 36.408, 72.820$ days*
Mid-transit time, $t_0 - 2,450,000$ (BJD <sub>TDB</sub> )	$7,322.51765 \pm 0.00025$	$7,362.72618 \pm 0.00033$	$7294.7741 \pm 0.0013^\dagger$
Transit depth, $(R_p/R_*)^2$	$0.754\% \pm 0.025\%$	$0.672\% \pm 0.042\%$	$0.826\% \pm 0.073\%^\dagger$
Transit impact parameter, $b$	$(0.21 \pm 0.14)R_*$	$(0.25 \pm 0.15)R_*$	$(0.24 \pm 0.15)R_*^\dagger$
Transit duration, $W$	$36.12 \pm 0.46$ min	$41.78 \pm 0.81$ min	$83.3 \pm 2.5$ min <sup>†</sup>
Orbital inclination, $i$	$89.41 \pm 0.41$ deg	$89.50 \pm 0.31$ deg	$89.87 \pm 0.10$ deg <sup>†</sup>
Orbital eccentricity, $e$	0 (fixed)	0 (fixed)	0 (fixed)
Radius, $R_p$	$(1.113 \pm 0.044)R_\oplus$	$(1.049 \pm 0.050)R_\oplus$	$(1.168 \pm 0.068)R_\oplus^\dagger$
Scale parameter, $a/R_*$	$20.45^{+0.43}_{-0.81}$	$28.0^{+0.6}_{-1.1}$	$41\text{--}271^\ddagger$
Semi-major axis, $a$	$0.01111 \pm 0.00040$ AU	$0.01522 \pm 0.00055$ AU	$0.022\text{--}0.146$ AU <sup>‡</sup>
Irradiation, $S_p$	$(4.25 \pm 0.38)S_\oplus$	$(2.26 \pm 0.21)S_\oplus$	$(0.02\text{--}1.0)S_\oplus^\ddagger$
Equilibrium temperature, $T_{\text{eq}}$			
with Bond albedo of 0.00	$400 \pm 9$ K	$342 \pm 8$ K	$110\text{--}280$ K <sup>‡</sup>
with Bond albedo of 0.75	$285 \pm 7$ K	$242 \pm 6$ K	$75\text{--}200$ K <sup>‡</sup>

The values and  $1\sigma$  errors given for the planetary parameters and for the stellar mass ( $M_*$ ), radius ( $R_*$ ), density ( $\rho_*$ ), and effective temperature ( $T_{\text{eff}}$ ) were deduced from a global analysis of the photometric data, including *a priori* knowledge of the stellar properties (see Methods). BJD<sub>TDB</sub>, barycentric Julian date in the barycentric dynamical time standard.  $L_\odot$ ,  $M_\odot$ ,  $R_\odot$  and  $\rho_\odot$  are, respectively, the luminosity, mass, radius and density of the Sun.  $R_p$  and  $S_p$  are, respectively, the radius and irradiation of the planet;  $R_\oplus$  and  $S_\oplus$  are, respectively, the radius and irradiation of Earth.

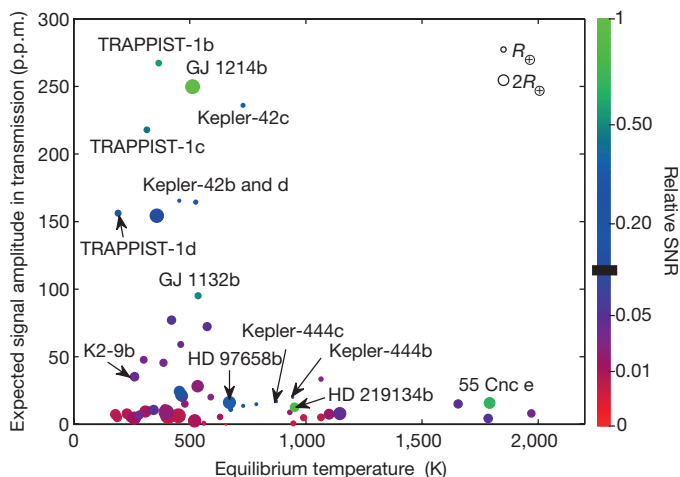
\*These are the potential orbital periods of TRAPPIST-1d, derived from non-continuous observations. The value in bold type is the most likely value for the period, as derived from the shape of the transits.

<sup>†</sup>Values calculated on the basis that  $P = 18.20175 \pm 0.00045$  days.

<sup>‡</sup>The ranges allowed by the set of possible periods.

and the past and present amplitudes of atmospheric replenishment mechanisms (impacts and volcanism). Fortunately, the TRAPPIST-1 planets are particularly well suited for detailed atmospheric characterization—notably by transmission spectroscopy (Fig. 3)—because transit signals are inversely proportional to the square of the host-star radius, the latter being only around 12% of that of the Sun for TRAPPIST-1. Data obtained by the Hubble Space Telescope should

provide initial constraints on the extent and composition of the planets' atmospheres. The next generation of observatories will then allow far more in-depth exploration of the atmospheric properties. In particular, data from the James Webb Space Telescope should yield strong constraints on atmospheric temperatures and on the abundances of molecules with large absorption bands including several potential biomarkers such as water, carbon dioxide, methane and ozone.



**Figure 3 | Potential for characterizing the atmospheres of known transiting sub-Neptune-sized exoplanets.** The signal being transmitted from each planet is estimated in parts per million (p.p.m.) and for transparent water-dominated atmospheres with a mean molecular weight,  $\mu$ , of 19. The signal-to-noise ratio (SNR) in transmission (normalized to that of GJ 1214b under the same atmospheric assumptions) is also calculated. The estimated signal and SNR are plotted against equilibrium temperatures, assuming a Bond albedo of 0.3. The black horizontal bar indicates the SNR that will require 200 (or 500) [or 1,000] hours of in-transit observations with the James Webb Space Telescope to yield a planet's atmospheric temperature with a relative uncertainty below 15% and with abundances within a factor of four in the case of a  $\text{H}_2\text{O}$  (or  $\text{N}_2$ ) [or  $\text{CO}_2$ ]-dominated atmosphere ( $\mu = 19$  (or 28) [or 39]). Only the exoplanets with a measured radius equal to or smaller than that of GJ 1214b are included in the figure. The size of the circular symbol for each planet is proportional to the planet's physical size. For illustration, symbols for planets of one ( $R_\oplus$ ) and two ( $2R_\oplus$ ) Earth radii are shown at the top right.



**Online Content** Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

**Received 11 January; accepted 18 February 2016.**

**Published online 2 May 2016.**

- Kirkpatrick, J. D., Henry, T. J. & Simon, D. A. The solar neighborhood. II. The first list of dwarfs with spectral types of M7 and cooler. *Astron. J.* **109**, 797–807 (1995).
- Cantrell, J. R., Henry, T. J. & White, R. J. The solar neighborhood XXIX: the habitable real estate of our nearest stellar neighbours. *Astron. J.* **146**, 99 (2013).
- Andrews, S. M., Wilner, D. J., Hugues, A. M., Qi, C. & Dullemond, C. P. Protoplanetary disk structures in Ophiuchus. II. Extension to fainter sources. *Astrophys. J.* **723**, 1241–1254 (2010).
- Liu, Y., Joergens, V., Bayo, A., Nielbock, M. & Wang, H. A homogeneous analysis of disk around brown dwarfs. *Astron. Astrophys.* **582**, A22 (2015).
- Payne, M. J. & Lodato, G. The potential for Earth-mass planet formation around brown dwarfs. *Mon. Not. R. Astron. Soc.* **381**, 1597–1606 (2007).
- Raymond, S. N., Scalo, J. & Meadows, V. S. A decreased probability of habitable planet formation around low-mass stars. *Astrophys. J.* **669**, 606–614 (2007).
- Montgomery, R. & Laughlin, G. Formation and detection of Earth-mass planets around low mass stars. *Icarus* **202**, 1–11 (2009).
- Kopparapu, R. K. *et al.* Habitable zones around main-sequence stars: new estimates. *Astrophys. J.* **765**, 131 (2013).
- Gillon, M. *et al.* TRAPPIST: a robotic telescope dedicated to the study of planetary systems. *EPJ Web Conf.* **11**, 06002 (2011).
- Gillon, M., Jehin, E., Fumel, A., Magain, P. & Queloz, D. TRAPPIST-UCDTS: a prototype search for habitable planets transiting ultra-cool stars. *EPJ Web Conf.* **47**, 03001 (2013).
- Liebert, J. & Gizis, J. E. RI photometry of 2MASS-selected late M and L dwarfs. *Publ. Astron. Soc. Pacif.* **118**, 659–670 (2006).
- Costa, E. *et al.* The solar neighborhood. XVI. Parallaxes from CTIOPI: final results from the 1.5m telescope program. *Astron. J.* **132**, 1234–1247 (2006).
- Filippazzo, J. C. *et al.* Fundamental parameters and spectral energy distributions of young and field age objects with masses spanning the stellar to planetary regime. *Astrophys. J.* **810**, 158 (2015).
- Gillon, M. *et al.* The TRAPPIST survey of southern transiting planets. I. Thirty eclipses of the ultra-short period planet WASP-43 b. *Astron. Astrophys.* **542**, A4 (2012).
- Reiners, A. & Basri, G. A volume-limited sample of 63 M7–M9.5 dwarfs. II. Activity, magnetism, and the fade of the rotation-dominated dynamo. *Astrophys. J.* **710**, 924–935 (2010).
- Hosey, A. D. *et al.* The solar neighbourhood. XXXVI. The long-term photometric variability of nearby red dwarfs in the VRI optical bands. *Astron. J.* **150**, 6 (2015).
- Yu, L. *et al.* Tests of the planetary hypothesis for PTF08–8695b. *Astrophys. J.* **812**, 48 (2015).
- Stelzer, B., Marino, A., Micela, G., López-Santiago, J. & Liefke, C. The UV and X-ray activity of the M dwarfs within 10 pc of the Sun. *Mon. Not. R. Astron. Soc.* **431**, 2063–2079 (2013).
- Lopez, E. D., Fortney, J. J. & Miller, N. How thermal evolution and mass-loss sculpt populations of super-Earths and sub-Neptunes: application to the Kepler-11 system and beyond. *Astrophys. J.* **761**, 59 (2012).
- Rogers, L. A. Most 1.6 Earth-radius planets are not rocky. *Astrophys. J.* **801**, 41 (2015).
- Wolfgang, A. & Lopez, E. How rocky are they? The composition distribution of Kepler's sub-Neptune planet candidates within 0.15 AU. *Astrophys. J.* **806**, 183 (2015).
- Seager, S., Kuchner, M., Hier-Majumder, C. A. & Militzer, B. Mass-radius relationships for solid exoplanets. *Astrophys. J.* **669**, 1279–1297 (2007).
- Holman, M. J. & Murray, N. W. The use of transit timing to detect terrestrial-mass extrasolar planets. *Science* **307**, 1288–1291 (2005).

- de Wit, J. & Seager, S. Constraining exoplanet mass from transmission spectroscopy. *Science* **342**, 1473–1477 (2013).
- Kasting, J. F., Whitmire, D. P. & Reynolds, R. T. Habitable zones around main-sequence stars. *Icarus* **101**, 108–128 (1993).
- Leconte, J. *et al.* 3D climate modelling of close-in land planets: circulation patterns, climate moist instability, and habitability. *Astron. Astrophys.* **554**, A69 (2013).
- Menou, K. Water-trapped world. *Astrophys. J.* **774**, 51 (2013).
- Driscoll, P. E. & Barnes, R. Tidal heating of Earth-like exoplanets around M stars: thermal, magnetic, and orbital evolutions. *Astrobiology* **15**, 739–760 (2015).
- France, K. *et al.* The ultraviolet radiation environment around M dwarf exoplanet host stars. *Astrophys. J.* **763**, 149 (2013).
- Tian, F. & Ida, S. Water contents of Earth-mass planets around M-dwarfs. *Nature Geosci.* **8**, 177–180 (2015).

**Acknowledgements** TRAPPIST is funded by the Belgian Fund for Scientific Research (FRS–FNRS) under grant FRFC 2.5.594.09.F, with the participation of the Swiss Fund for Scientific Research. The research leading to our results was funded in part by the European Research Council (ERC) under the FP/2007–2013 ERC grant 336480, and through an Action de Recherche Concertée (ARC) grant financed by the Wallonia-Brussels Federation. Our work was also supported in part by NASA under contract NNX15AI75G. UKIRT is supported by NASA and operated under an agreement among the University of Hawaii, the University of Arizona, and Lockheed Martin Advanced Technology Center; operations are enabled through the cooperation of the East Asian Observatory. The facilities at the Indian Astronomical Observatory (IAO) and the Consortium for Research Excellence, Support and Training (CREST) are operated by the Indian Institute of Astrophysics, Bangalore. M.G., E.J. and V.V.G. are FRS–FNRS research associates. L.D. and C.O. are FRS–FNRS PhD students. We thank V. Mégevand, the ASTELCO telescope team, S. Sohy, V. Chantry, and A. Fumel for their contributions to the TRAPPIST project; the Infrared Telescope Facility (IRTF) operators B. Cabreira and D. Griep for assistance with the SpeX observations; UKIRT staff scientists W. Varricatt & T. Kerr, telescope operators S. Benigni, E. Moore and T. Carroll, and Cambridge Astronomy Survey Unit (CASU) scientists G. Madsen and M. Irwin for assistance with UKIRT observations; the European Southern Observatory (ESO) astronomers A. Smette and G. Hau for providing us with the best possible VLT data; and the staff of IAO (in Hanle) and CREST (in Hosakote) for making observations with the HCT possible. Ad.B. and D.B.G. are visiting astronomers at the IRTF, which is operated by the University of Hawaii under Cooperative Agreement NNX-08AE38A with NASA's Science Mission Directorate, Planetary Astronomy Program.

**Author Contributions** The TRAPPIST team (M.G., E.J., L.D., A.B., C.O. and P.M.) discovered the planets. M.G. leads the exoplanet program of TRAPPIST, set up and organized the ultracool-dwarf transit survey, planned and analysed part of the observations, led their scientific exploitation, and wrote most of the manuscript. E.J. manages the maintenance and operations of the TRAPPIST telescope. S.M.L. obtained the director's discretionary time on UKIRT, and managed, with E.J., the preparation of the UKIRT observations. L.D. and C.O. scheduled and carried out some of the TRAPPIST observations. L.D. and A.B. analysed some photometric observations. J.d.W. led the study of the amenability of the planets for detailed atmospheric characterization. V.V.G. checked the physical parameters of the star. A.J.B. checked the spectral type of the star and determined its metallicity. B.-O.D. took charge of the dynamical simulations. D.B.G. acquired the SpeX spectra. D.K.S. gathered the HCT observations. S.M.L., A.H.M.J.T., P.M. and D.Q. helped to write the manuscript. A.H.M.J.T. prepared most of the figures.

**Author Information** Reprints and permissions information is available at [www.nature.com/reprints](http://www.nature.com/reprints). The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to M.G. ([michael.gillon@ulg.ac.be](mailto:michael.gillon@ulg.ac.be)).

## METHODS

**Spectral type, parallax and age of the star.** TRAPPIST-1 = 2MASS J23062928 – 0502285 was discovered in 2000 by a search for nearby ultracool dwarfs according to photometric criteria<sup>31</sup>, and identified as a high proper-motion (right ascension component  $\mu_{\alpha} = 0.89''$ , declination component  $\mu_{\delta} = -0.42''$ ), moderately active (logarithm of  $H_{\alpha}$  to bolometric luminosity ratio  $\log L_{H\alpha}/L_{\text{bol}} = -4.61$ ), M7.5 dwarf at approximately 11 parsecs from Earth. Subsequent findings converged on a spectral type of  $M8.0 \pm 0.5$  (refs 11, 32), while confirming a moderate level of activity typical of stars of similar spectral type in the vicinity of the Sun<sup>15,33,34</sup>. The spectral classification was checked by comparing a low-resolution ( $R \approx 150$ ) near-infrared spectrum of the star<sup>13</sup>—obtained with the SpeX spectrograph<sup>35</sup> mounted on the 3-metre NASA Infrared Telescope Facility—with several spectral-type standards; the spectrum of TRAPPIST-1 best fit that of the M8-type standard LHS 132 (Extended Data Fig. 3a). The Cerro Tololo Interamerican Observatory Parallax Investigation (CTIOPI) project reported the star's trigonometric parallax to be  $\pi = 82.6 \pm 2.6$  mas (ref. 12), which translates to a distance of  $12.1 \pm 0.4$  parsecs. High-resolution optical spectroscopy failed to detect significant absorption at the 6,708 Å lithium line<sup>36</sup>, suggesting that the object is not a very young brown dwarf, but rather a very-low-mass main-sequence star. This is in agreement with its thick disk kinematics<sup>36</sup>, its relatively slow rotation (projected rotational velocity  $v \sin i = 6 \pm 2 \text{ km s}^{-1}$ )<sup>15</sup>, its moderate activity, and its reported photometric stability<sup>16</sup>, all of which point to an age of at least 500 Myr (ref. 13).

**Metallicity of the star.** We obtained new, near-infrared (0.9–2.5 µm) spectrographic data for TRAPPIST-1 with the SpeX spectrograph on the night of 18 November 2015 (universal time), during clear conditions and  $0.8''$  seeing at K-band. We used the cross-dispersed mode and  $0.3'' \times 15''$  slit, aligned at the parallactic angle, to acquire moderate-resolution data ( $\lambda/\Delta\lambda \approx 2,000$ ) with a dispersion of 3.6 Å per pixel, covering the spectral range 0.9–2.5 µm in seven orders. Ten exposures each of 300 seconds were obtained over an air mass ranging from 1.14 to 1.17, followed by observations of the A0V star 67 Aqr ( $V = 6.41$ ) at an air mass of 1.19 for telluric and flux calibration, as well as internal lamp exposures. Data were reduced using the SpeXtool package version 4.04 (refs 37, 38). The reduced spectrum has a median signal-to-noise ratio of 300 in the 2.17–2.35-µm region (see Extended Data Fig. 3b; the metallicity-sensitive atomic features of Na I (2.206 µm, 2.209 µm) and Ca I (2.261 µm, 2.263 µm, 2.266 µm) are labelled). We measured the equivalent widths of these features and the  $H_2O$ –K2 index (defined in ref. 39), and used the mid- and late-M-dwarf metallicity calibration of ref. 40 to determine  $[\text{Fe}/\text{H}] = 0.04 \pm 0.02$  (measurement)  $\pm 0.07$  (systematic) for TRAPPIST-1. The quadratic sum of the two errors resulted in our final value for  $[\text{Fe}/\text{H}]$  of  $0.04 \pm 0.08$ .

**Basic parameters of the star.** A recent study<sup>13</sup> derived a luminosity for TRAPPIST-1 of  $(0.000525 \pm 0.000036)L_{\odot}$  (where  $L_{\odot}$  is the luminosity of the Sun), using as input data the trigonometric parallax and VRI magnitudes as measured by the Cerro Tololo Interamerican Observatory Parallax Investigation (CTIOPI) project<sup>12</sup>, the 2MASS JHK magnitudes<sup>41</sup>, the Wide-Field Infrared Survey Explorer (WISE) W123 magnitudes<sup>42</sup>, an optical spectrum measured with the Kitt Peak National Observatory Ritchey–Chrétien spectrograph<sup>43</sup>, and a near-infrared spectrum measured by SpeX/Prism. Using this luminosity and an age constraint of  $>500$  Myr, the authors of ref. 13 derived (from evolutionary-model isochrones and the Stefan–Boltzmann law) the following values for the mass, radius and effective temperature of TRAPPIST-1:  $M_{\star} = (0.082 \pm 0.009)M_{\odot}$ ,  $R_{\star} = (0.116 \pm 0.004)R_{\odot}$ , and  $T_{\text{eff}} = 2,557 \pm 64$  K. To account for the uncertainties coming from the assumptions and details of the evolutionary models, we carried out a new determination of these three basic parameters, using recent solar metallicity evolutionary-model isochrones that consistently couple atmosphere and interior structures<sup>44</sup>. We obtained  $M_{\star} = 0.089M_{\odot}$ ,  $R_{\star} = 0.112R_{\odot}$ , and  $T_{\text{eff}} = 2,615$  K. We then added the difference between the two determinations quadratically to the errors of ref. 13, adopting finally  $M_{\star} = (0.082 \pm 0.011)M_{\odot}$ ,  $R_{\star} = (0.116 \pm 0.006)R_{\odot}$ , and  $T_{\text{eff}} = 2,555 \pm 85$  K. We took the normal distributions corresponding to these values and errors as prior probability distribution functions in the Bayesian analysis of our photometric data (see below).

**Possible binary nature of the star.** High-resolution imaging from the ground<sup>45–47</sup> and from space with the Hubble Space Telescope<sup>48</sup> discarded the existence of a companion down to an angular distance of  $0.1''$ , corresponding to a projected physical distance of 1.2 au at 12 parsecs, and in good agreement with the reported stability of the radial velocity of the star at the  $\sim 10 \text{ m s}^{-1}$  level over one week<sup>49</sup> and at the  $\sim 150 \text{ m s}^{-1}$  level over about ten weeks<sup>50</sup>. We performed spectral binary template fitting<sup>51</sup> to the IRTF/SpeX spectroscopy, and statistically reject the presence of an L- or T-type brown-dwarf companion that would be visible in a blended-light spectrum. TRAPPIST-1 can thus, in all probability, be considered to be an isolated star.

**Upper magnitude limits on a background eclipsing binary.** We measured the J2000 equatorial coordinates of TRAPPIST-1 in the 2015 TRAPPIST images, using 29 stars from the UCAC2 catalogue<sup>52</sup> and the Pulkovo Observatory

Izmcdd astrometric software<sup>53</sup>. We obtained coordinates of right ascension (RA) = 23 h 06 min 30.34 s and declination (dec.) =  $-05^{\circ}02'36.44''$ . Owing to the high proper motion of TRAPPIST-1 ( $\sim 1''$  per year), we could assess the possible presence of a background object by examining this exact position in several previous images taken from the POSS (1953; ref. 54) and 2MASS (1998; ref. 41) image catalogues. We detected no possible additional source at this position in any of these images. The faintest stars detected at other positions in the 2MASS images have J-band magnitudes of  $\sim 17$ . We adopt this value as an absolute lower threshold for the J-band magnitude of a background source blended with TRAPPIST-1 in our TRAPPIST 2015 images. TRAPPIST-1 has a J-band magnitude of 11.35 (ref. 42), and the achromaticity of the transits of TRAPPIST-1b as observed from 0.85 µm to 2.09 µm means that, if the transits originated from a background eclipsing binary (BEB), then that BEB would have to be a very red object with a spectral type similar to that of TRAPPIST-1. Combining these two facts, the BEB scenario would require an unphysical eclipse depth of more than 100% in the photometric bands probed by our observations to match the  $\sim 0.8\%$  depths measured after dilution by the light from TRAPPIST-1. We thus firmly discard the BEB scenario.

**Photometric observations and analysis.** The TRAPPIST<sup>8,55</sup> observations in which the transits were detected consisted of 12,295 exposures, each of 55 seconds, gathered with a thermoelectrically cooled 2Kx2K CCD camera (field of view of  $22' \times 22'$ ; pixel scale of  $0.65''$ ). Most of the observations were obtained through an I + z filter with a transmittance greater than 90% from 750 nm to beyond 1,100 nm—the effective bandpass in this spectral range being defined by the response of the CCD. On the basis of the spectral efficiency model for TRAPPIST and an optical spectrum of a spectroscopic standard M8V star (VB10), we compute an effective wavelength of  $885 \pm 5$  nm for these observations. For the nights of 20 November and 19 December 2015, the target was close to the full Moon and the observations were performed in the Sloan z' filter to minimize the background. After a standard pre-reduction (bias, dark, flat-field correction), the TRAPPIST automatic pipeline extracted the stellar fluxes from the images using the DAOPHOT aperture photometry software<sup>56</sup> for eight different apertures. A careful selection of both the photometric aperture size and the stable comparison stars was then performed manually to obtain the most accurate differential light curves of the target.

Photometric follow-up observations were performed with the HAWK-I near-infrared imager<sup>57</sup> on the European Southern Observatory (ESO) 8-metre Very Large Telescope (Chile), with the HFOSC optical spectro-imager ([http://www.iap.res.in/iao\\_hfosc](http://www.iap.res.in/iao_hfosc)) on the 2-metre Himalayan Chandra Telescope (India), and with the WFCAM<sup>58</sup> near-infrared camera located at the prime focus of the 3.8-metre UKIRT telescope (Hawaii).

The VLT/HAWK-I observations of a transit of planet TRAPPIST-1b were performed during the night of 8 November 2015. HAWK-I is composed of four Hawaii 2RG  $2,048 \times 2,048$  pixel detectors (pixel scale =  $0.106''$ ). Its total field of view on the sky is  $7.5' \times 7.5'$ . The transit was observed through the narrowband filter NB2090 ( $\lambda = 2.095 \mu\text{m}$ , width =  $0.020 \mu\text{m}$ ). 185 exposures, composed of 17 integrations of 1.7 seconds each, were acquired during the run in 'stare' mode—that is, without applying a jitter pattern. After standard calibration of the images, stellar flux measurement was performed by aperture photometry<sup>14</sup>.

The HCT/HFOSC observations of a transit of TRAPPIST-1b were performed on 18 November 2015. The imager in the HFOSC CCD detector is an array of  $2,048 \times 2,048$  pixels, corresponding to a field of view of  $10' \times 10'$  on-sky (pixel scale =  $0.3''$ ). The observations consisted of 104 exposures, each of 20 seconds, taken in stare mode and in the I filter, centred on the expected transit time. After a standard calibration of these images and their photometric reduction with DAOPHOT, differential photometry was performed. We estimate the effective wavelength of these observations to be  $840 \pm 20$  nm, given the spectral response of HFOSC and an optical spectrum of the M8V standard star VB10.

The UKIRT/WFCAM observations of two transits of planet TRAPPIST-1b and one transit of planet TRAPPIST-1c consisted of three runs of 4 hours each, performed on 5, 6 and 8 December 2015 in the J-band. WFCAM is composed of four HgCdTe detectors of  $2,048 \times 2,048$  pixels each, with a pixel scale of  $0.4''$ , resulting in a field of view of  $13.65' \times 13.65'$  for each detector. On 5 December 2015, 1,365 exposures composed of three integrations of 2 seconds each were performed in stare mode. For the runs on 6 and 8 December 2015, respectively, 1,181 and 1,142 exposures composed of five one-second exposures were performed, again in stare mode and using the same pointing as on 5 December 2015. Differential aperture photometry was performed with DAOPHOT on all calibrated images.

**Global analysis of the photometry.** We inferred the parameters of the three detected planets transiting TRAPPIST-1 from analysis of their transit light curves (Extended Data Fig. 1 and Extended Data Table 1) with an adaptive Markov chain Monte Carlo (MCMC) code<sup>14</sup>. We converted each universal time (UT) of mid-exposure to the BJD<sub>TDB</sub> time system<sup>59</sup>. The model assumed for each light curve was composed of the eclipse model of ref. 60, multiplied by a baseline model, aiming to



represent the other astrophysical and instrumental mechanisms able to produce photometric variations. Assuming the same baseline model for all light curves, and minimizing the Bayesian information criterion (BIC)<sup>61</sup>, we selected a second-order time polynomial as a baseline model to represent the curvature of the light curves due to the differential extinction and the low-frequency variability of the star, and added an instrumental model composed of a second-order polynomial function of the positions and widths of the stellar images.

Stellar metallicity, effective temperature, mass and radius were four free parameters in the MCMC for which prior probability distribution functions (PDFs) were selected as input. Here, the normal distributions  $N(0.04, 0.08^2)$  dex,  $N(2,555, 85^2)$  K,  $N(0.082, 0.011^2)M_{\odot}$ , and  $N(0.116, 0.006^2)R_{\odot}$  were assumed on the basis of *a priori* knowledge of the stellar properties (see the section on 'Basic parameters of the star'). Circular orbits were assumed for all transiting objects. For each of them, the additional free parameters in the MCMC included: (1) the transit depth, dF, defined as  $(R_p/R_*)^2$ , with  $R_p$  and  $R_*$  being the planetary and stellar radii, respectively; (2) the transit impact parameter  $b = a \cos i / R_*$ , with  $a$  and  $i$  being the planet's semi-major axis and orbital inclination, respectively; (3) the orbital period  $P$ ; (4) the transit width  $W$  defined as  $(P \times R_p / R_*)^2 - b^2)^{1/2} / \pi$ ; and (5) the mid-transit time (time of inferior conjunction)  $T_0$ . Uniform prior distributions were assumed for each of these free parameters. At each step of the MCMC, values for  $R_p$ ,  $a$  and  $i$ , were computed from the values for the transit and stellar parameters; values were also computed for the irradiation of the planet in Earth units and for its equilibrium temperatures, assuming Bond albedos of 0 and 0.75, respectively. A quadratic limb-darkening law<sup>60</sup> was assumed for the star. For each bandpass, values and errors for the limb-darkening coefficients  $u_1$  and  $u_2$  were derived from the tables in ref. 62 (Extended Data Table 2), and the corresponding normal distributions were used as prior PDFs in the MCMC.  $u_1$  and  $u_2$  were free parameters under the control of these PDFs in the MCMC.

We divided our analysis into three phases. The first phase focused on the two inner planets, for which the period is firmly determined. A circular orbit was assumed for both planets. All transit light curves of the two planets were used as input data for this first phase, except the TRAPPIST light curve of 11 December 2015, for which the transit of planet TRAPPIST-1c is blended with a transit of planet TRAPPIST-1d. A preliminary MCMC analysis composed of one chain of 50,000 steps was first performed to estimate the need to rescale the photometric errors<sup>14</sup>. Then a longer MCMC analysis was performed, composed of five chains of 100,000 steps, whose convergence was checked using the statistical test of ref. 63. The parameters derived from this analysis for the star and its two inner planets are shown in Table 1. We performed a similar analysis assuming a uniform prior PDF for the stellar radius to derive the value of the stellar density constrained only by the transit photometry<sup>64</sup>. It resulted in a stellar density of  $49.3^{+4.1}_{-8.3} \rho_{\odot}$ , in excellent agreement with the density of  $(55.3 \pm 12.1) \rho_{\odot}$  derived from the *a priori* knowledge of the star, thus bringing a further validation of the planetary origin of the transit signals.

In the second phase of our analysis, we performed 11 global MCMC analyses of all transit light curves, each of them consisting of one chain of 50,000 steps and corresponding to one of the possible values of the period of TRAPPIST-1d (see Table 1) for which a circular orbit was assumed. We then repeated the 11 analyses under the assumption of an eccentric orbit for TRAPPIST-1d. We used the medians of the BIC posterior distributions to compare the relative posterior probability of each orbital model through the formula  $P1/P2 = e^{(\text{BIC}_2 - \text{BIC}_1)/2}$ . The resulting relative probabilities are given in Extended Data Table 3. The table shows that our data favour (with a relative probability of >10%) a circular orbit and an orbital period of between 10.4 and 36.4 days—the most likely period being 18.4 days.

In the final phase, we performed individual analyses of the light curves to measure the mid-eclipse time of each transit to support future TTV studies of the system<sup>23,65</sup>. The resulting timings are shown in Extended Data Table 4. They do not reveal any significant TTV signal, which is not surprising given the amplitude of the expected periodicity departures (see below) combined with the limited timing precision of the TRAPPIST photometry.

Extended Data Figs 1 and 2 show the raw and de-trended light curves, respectively; for each of these, the best-fit eclipse plus baseline model is overplotted. The phased-time de-trended light curves are shown for each planet and bandpass in Fig. 1.

**Photometric variability of the star.** We used the TRAPPIST data set to assess the photometric variability of the star at about 900 nm. On a timescale of a few hours—corresponding to the typical duration of our observing runs—the star appears to be relatively stable, except for the transits and for four sharp, low-amplitude increases in brightness (of one to a few per cent) that are followed by exponential-type decreases to normal levels within 10–15 minutes (Extended Data Fig. 4), which we attribute to flares<sup>66</sup>. The low amplitude and inferred low frequency ( $1/60 \text{ h}^{-1}$ ) of these flares is consistent with the reported low level of activity of the star<sup>15,33,34</sup>, strengthening the inference that the system is not young.

To assess the lower-frequency variability of TRAPPIST-1, we built its global differential light curve in the I + z filter, using four stable stars of similar brightness in the TRAPPIST images as comparison stars. We filtered out the flares, transits, and measurements taken in cloudy conditions to create the resulting light curve, consisting of 12,081 photometric measurements. Extended Data Fig. 5a compares this light curve to that for the comparison star 2MASS J23063445 – 0507511. It clearly shows some variability at the level of a few per cent, which is consistent with previous photometric results obtained in the I-band<sup>16</sup>. A Lomb–Scargle (LS) periodogram<sup>67</sup> analysis of the light curve—from which low-frequency variations and differential extinction have been filtered out by division of the best-fit fourth-order polynomial in time and air mass—reveals a power excess with a period of 1.4 days (see Extended Data Fig. 5b). Cutting the light curves in two, and in four in a second test, and performing a LS analysis of each fraction revealed a power excess at about 1.4 days for all of them, supporting a genuine periodic signal of astrophysical origin. Associating it with the stellar rotation period, the resulting equatorial rotation speed of  $4.2 \text{ km s}^{-1}$  (assuming  $R_* = 0.117 R_{\odot}$ ) is consistent with the literature measurement<sup>15</sup> for  $v \sin i$  of  $6 \pm 2 \text{ km s}^{-1}$ , making this association physically meaningful. Given the scatter of the peak values obtained in the LS analyses of the light-curve fractions, we estimate the error bar on the rotation period of 1.40 days to be 0.05 days. In summary, the photometric variability of the star seems to be dominated by the rotation and evolution of photospheric inhomogeneities (spots) combined with rare flares.

**Dynamics of the system.** We computed the tidal circularization timescales<sup>68</sup>

of the three planets according to  $t_{\text{circ}} = \frac{2PQ}{63\pi} \frac{M_p}{M_*} \left( \frac{a}{R_p} \right)^5$ , assuming planetary masses,

$M_p$ , ranging from 0.45 Earth masses (pure ice composition) to 3 Earth masses (pure iron composition)<sup>22</sup> and a tidal quality factor<sup>69</sup>,  $Q$ , of 100, corresponding to the maximum value derived for terrestrial planets and satellites of the solar system<sup>69</sup>. For planets TRAPPIST-1b and -1c, the computed values range from 22 Myr to 145 Myr and from 177 Myr to 1.1 Gyr, respectively. Taking into account that the system is apparently not very young and that the orbits have weak mutual perturbations (as they are not close to any mean-motion resonance), our assumption of circular orbits for the two inner planets is reasonable. On the other hand, the same computations result in values ranging from a few to tens of billions of years for TRAPPIST-1d, making a significant orbital eccentricity possible from a tidal theory perspective. Nonetheless, a nearly circular orbit for this outer planet is still a reasonable hypothesis when considering the strong anticorrelation of orbital eccentricity and multiplicity of planets detected by radial velocities<sup>70</sup>, and is favoured by our global analysis of the transit photometry (see above).

We used the Mercury software package<sup>71</sup> to assess the dynamical stability of the system over 10,000 years for all possible periods of TRAPPIST-1d. Instabilities appeared in our simulations only for the unlikely scenarios of this planet on a significantly eccentric ( $e \geq 0.4$ ) 4.5-day or 5.2-day orbit.

To assess the potential of the TTV method<sup>23,65</sup> to measure the masses of the planets, we integrated the dynamical evolution of the system at high sampling over two years, assuming Earth masses for the three planets and an 18.4-day circular orbit for TRAPPIST-1d. These simulations resulted in TTV amplitudes of several tens of seconds, and led us to conclude that, with an intensive transit monitoring campaign—with instruments able to reach timing precisions of a few tens of seconds (for example, with VLT/HAWK-I or UKIRT/WFCAM; Extended Data Table 4)—it should be possible to constrain the planetary masses.

**Planets' suitability for atmospheric characterization.** We estimated the typical signal amplitude in transit transmission spectroscopy for all the transiting exoplanets with a size equal to or smaller than that of the mini-Neptune GJ 1214b (ref. 72). We computed this amplitude as  $2R_p h_{\text{eff}} / R_*^2$ , where  $R_p$  is the planetary radius,  $h_{\text{eff}}$  is the effective atmospheric height (that is, the extent of the atmospheric annulus), and  $R_*$  is the stellar radius. The effective atmospheric height is directly proportional to the atmospheric scale height,  $H = kT/\mu g$ , where  $k$  is Boltzmann's constant,  $T$  is the atmospheric temperature,  $\mu$  is the atmospheric mean molecular mass, and  $g$  is the surface gravity. The ratio  $h_{\text{eff}}/H$  for a transparent atmosphere<sup>24,73</sup> is typically between 6 and 10, and thus depends strongly on the presence of clouds and the spectral resolution and range covered. Our estimates (Fig. 2) are based on an  $h_{\text{eff}}/H$  ratio of 7 and the conservative assumption of a volatile-dominated atmosphere ( $\mu = 20$ ) with a Bond albedo of 0.3. All other parameters for the planets were derived from exoplanets.org<sup>74</sup>. As an illustration, the maximum transit depth variations projected under those assumptions for GJ 1214b are about 250 p.p.m., in agreement with independent simulations<sup>75</sup>.

For the same sample of planets, we also derived the typical SNRs in transit transmission spectroscopy from the ratio of our computed signal amplitudes over the square root of the flux (determined from the J-band magnitudes of the host stars). The SNRs of TRAPPIST-1's planets in transmission are expected to range between 0.22 and 0.55 times that of GJ 1214b under the same theoretical assumptions,



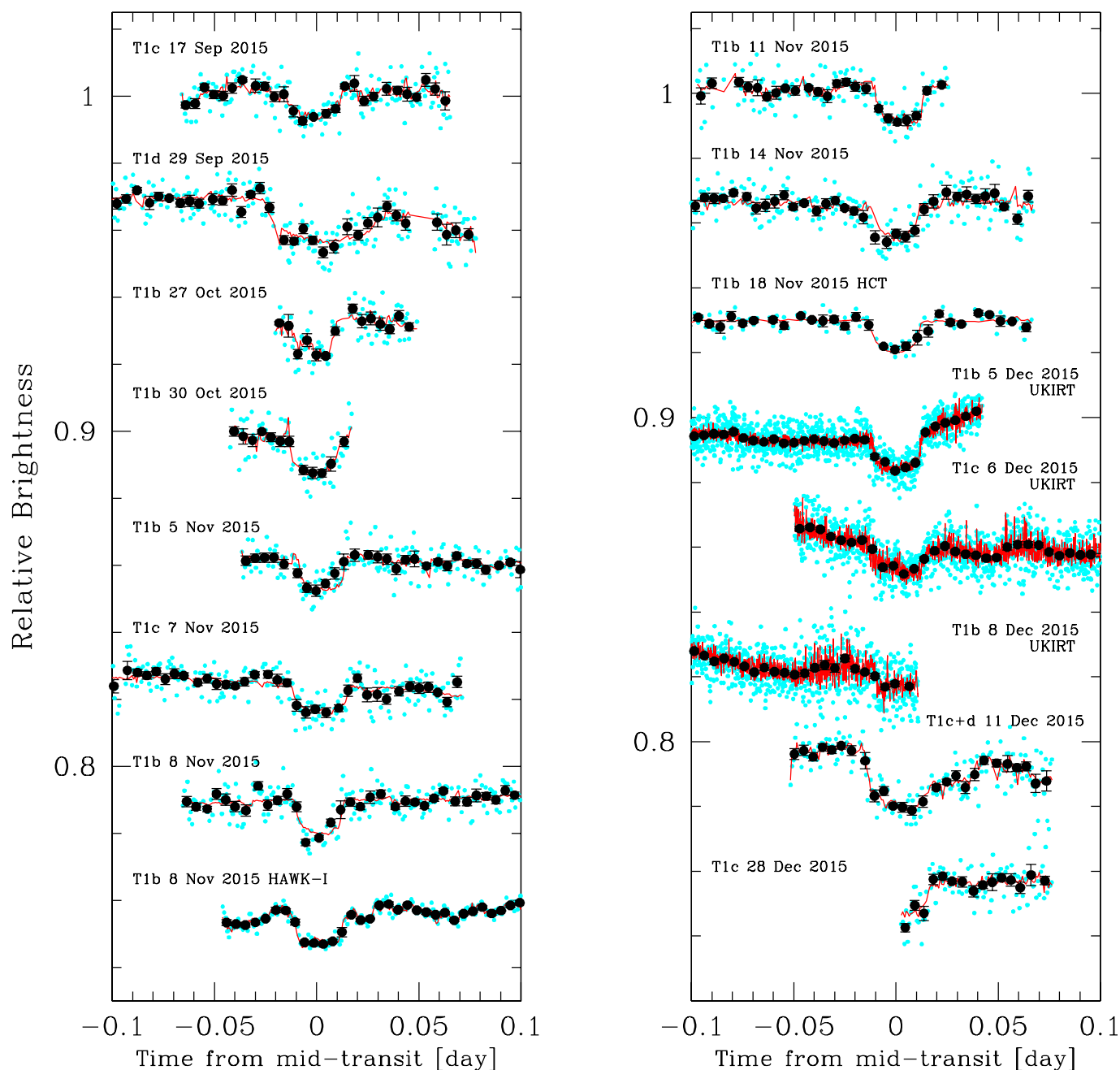
suggesting that these planets are well suited for atmospheric studies with HST/WFC3 similar to those previously targeting GJ 1214b (refs 75, 76).

Given published simulations for terrestrial planets<sup>24</sup>, we estimate that characterization of TRAPPIST-1b, -1c and -1d should require up to 70 hours, 90 hours and 270 hours, respectively, of in-transit observations with the James Webb Space Telescope (JWST), and should yield atmospheric temperatures with relative uncertainties below 15% and abundances within a factor of four. Assuming that the atmospheres of TRAPPIST-1's planets are not depleted and do not harbour a high-altitude cloud deck, JWST should, notably, yield constraints on the abundances of molecules with large absorption bands such as H<sub>2</sub>O, CO<sub>2</sub>, CH<sub>4</sub>, CO and O<sub>3</sub> if their abundances are at or greater than the 10-p.p.m. level.

We also assessed the potential of the cross-correlation technique<sup>77</sup> to constrain the atmospheric properties of the TRAPPIST-1 planets, following a published formalism<sup>78</sup>. We find that detecting O<sub>2</sub> in TRAPPIST-1's planets should require up to 80 transit observations with one of the next-generation, giant ground-based telescopes. Taking into account the limited fraction of transits visible at low air mass, such an endeavour could be reached in 5 to 15 years.

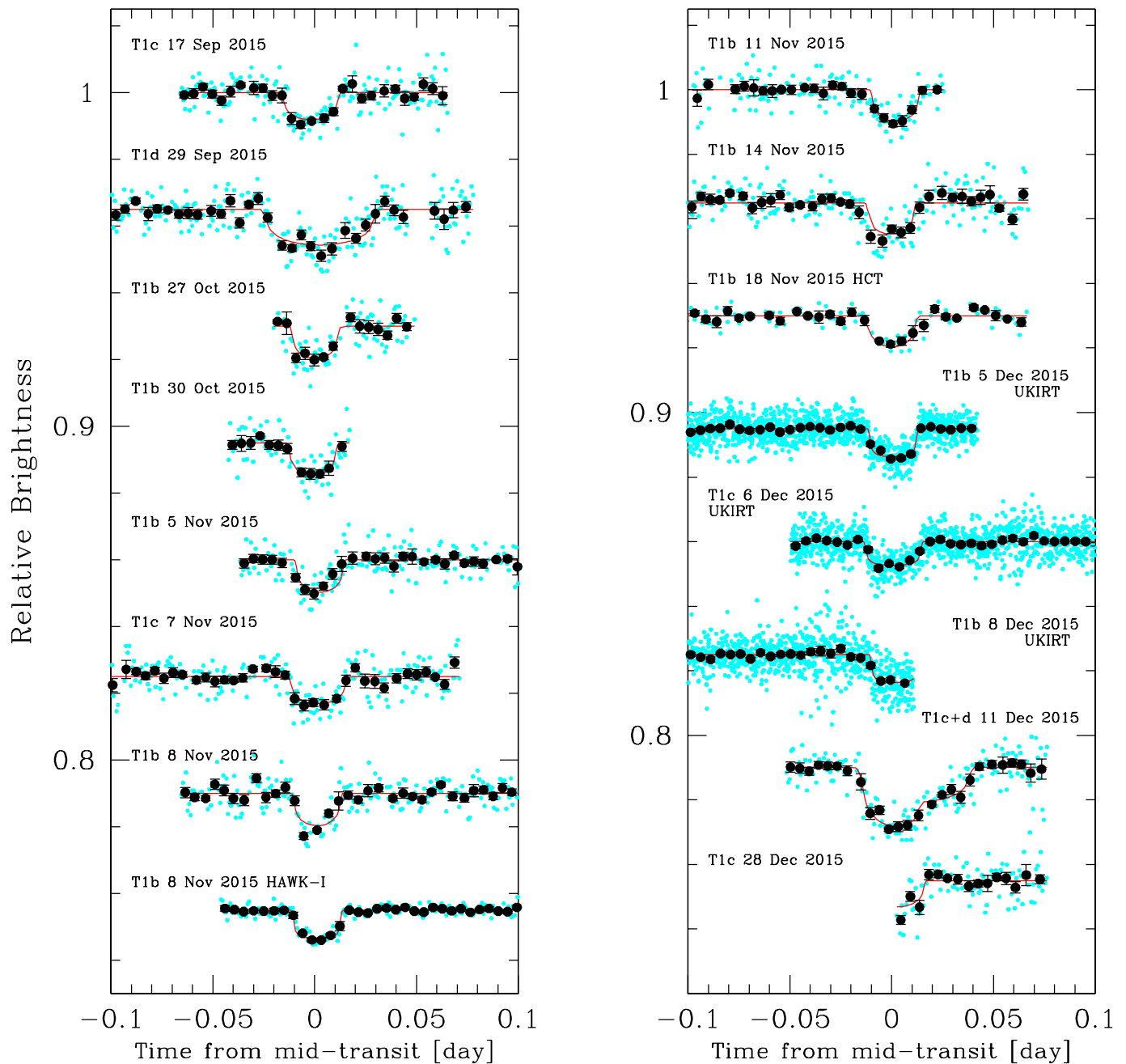
**Code availability.** Equivalent widths and H<sub>2</sub>O–K2 index measurements in the SpeX spectra were made using the IDL program created by A. Mann and distributed at <http://github.com/awmann/metal>. Conversion of the UT times for the photometric measurements to the BJD<sub>TDB</sub> system was performed using the online program created by J. Eastman and distributed at <http://astroutils.astronomy.ohio-state.edu/time/utc2bjd.html>. The Image Reduction and Analysis Facility (IRAF) software is distributed by the National Optical Astronomy Observatory, which is operated by the Association of Universities for Research in Astronomy, Inc., under cooperative agreement with the National Science Foundation. The MCMC software used to analyse the photometric data is a custom Fortran 90 code that can be obtained upon request.

31. Gizis, J. E. *et al.* New neighbours from 2MASS: activity and kinematics at the bottom of the main sequence. *Astron. J.* **120**, 1085–1099 (2000).
32. Bartlett, J. L. Knowing our neighbours: fundamental properties of nearby stars. *Publ. Astron. Soc. Pacif.* **119**, 828–829 (2007).
33. Schmidt, S. J., Cruz, K. L., Bongiorno, B. J., Liebert, J. & Reid, I. N. Activity and kinematics of ultracool dwarfs, including an amazing flare observation. *Astron. J.* **133**, 2258–2273 (2007).
34. Lee, K.-G., Berger, E. & Knapp, G. R. Short-term H $\alpha$  variability in M dwarfs. *Astrophys. J.* **708**, 1482–1491 (2010).
35. Rayner, J. T. *et al.* SpeX: a medium-resolution 0.8–5.5 micron spectrograph and imager for the NASA infrared telescope facility. *Publ. Astron. Soc. Pacif.* **115**, 362–382 (2003).
36. Reiners, A. & Basri, G. A volume-limited sample of 63 M7–M9.5 dwarfs. I. Space motion, kinematics age, and lithium. *Astrophys. J.* **705**, 1416–1424 (2009).
37. Vacca, W. D., Cushing, M. C. & Rayner, J. T. A method of correcting near-infrared spectra for telluric absorption. *Publ. Astron. Soc. Pacif.* **115**, 389–409 (2003).
38. Cushing, M. C., Vacca, W. D. & Rayner, J. T. Spextool: a spectral extraction package for SpeX, a 0.8–5.5 micron cross-dispersed spectrograph. *Publ. Astron. Soc. Pacif.* **116**, 362–376 (2004).
39. Rojas-Ayala, B., Covey, K. R., Muirhead, P. S. & Lloyd, J. P. Metallicity and temperature indicators in M dwarf K-band spectra: testing new and updated calibrations with observations of 133 solar neighbourhood M dwarfs. *Astrophys. J.* **748**, 93 (2012).
40. Mann, A. W. *et al.* Prospecting in ultracool dwarfs: measuring the metallicities of mid- and late-M dwarfs. *Astron. J.* **147**, 160 (2014).
41. Skrutskie, M. F., Meyer, M. R., Whalen, D. & Hamilton, C. The two micron all sky survey (2MASS). *Astron. J.* **131**, 1163–1183 (2006).
42. Cutri, R. M. *et al.* VizieR online data catalog II/311: WISE all-sky data release. <http://vizier.cfa.harvard.edu/viz-bin/VizieR?source=II/311> (2012).
43. Cruz, K. L. *et al.* Meeting the cool neighbours. IX. The luminosity function of M7–L8 ultracool dwarfs in the field. *Astron. J.* **133**, 439–467 (2007).
44. Baraffe, I., Homeier, D., Allard, F. & Chabrier, G. New evolutionary models for pre-main sequence and main sequence low-mass stars down to the hydrogen-burning limit. *Astron. Astrophys.* **577**, A42 (2015).
45. Siegler, N., Close, L. M., Mamajek, E. E. & Freed, M. An adaptive optics survey of M6.0–M7.5 stars: discovery of three very low mass binary system including two probable Hyades member. *Astrophys. J.* **598**, 1265–1276 (2003).
46. Siegler, N., Close, L. M., Cruz, K. L., Martín, E. L. & Reid, I. N. Discovery of two very low mass binaries: final results of an adaptive optics survey of nearby M6.0–M7.5 stars. *Astrophys. J.* **621**, 1023–1032 (2005).
47. Janson, M. *et al.* The AstraLux large M-dwarf multiplicity survey. *Astrophys. J.* **754**, 44 (2012).
48. Bouy, H. *et al.* Multiplicity of nearby free-floating ultracool dwarfs: a Hubble Space Telescope WFC2 search for companions. *Astron. J.* **126**, 1526–1554 (2003).
49. Barnes, J. R. *et al.* Precision radial velocities of 15 M5–M9 dwarfs. *Mon. Not. R. Astron. Soc.* **439**, 3094–3113 (2014).
50. Tanner, A. *et al.* Keck NIRSPEC radial velocity observations of late M-dwarfs. *Astrophys. J.* **203** (Suppl.), 10 (2012).
51. Burgasser, A. J. *et al.* WISE J072003.20-084651.2: an old and active M9.5 + T5 spectral binary 6 pc from the Sun. *Astron. J.* **149**, 104 (2015).
52. Zacharias, N. *et al.* The second US Naval Observatory CCD astrophotograph catalog (UCAC2). *Astron. J.* **127**, 3043–3059 (2004).
53. Izmailov, I. S. *et al.* Astrometric CCD observations of visual double stars at the Pulkovo Observatory. *Astron. Lett.* **36**, 349–354 (2010).
54. Minkowski, R. L. & Abell, G. O. in *Basic Astronomical Data: Stars and Stellar Systems* (ed. Strand, K. A.) 481–487 (Univ. Chicago Press, 1963).
55. Jehin, E. *et al.* TRAPPIST: TRAnsiting Planets and Planetesimals Small Telescope. *The Messenger* **145**, 2–6 (2011).
56. Stetson, P. B. DAOPHOT—a computer program for crowded-field stellar photometry. *Publ. Astron. Soc. Pacif.* **99**, 191–222 (1987).
57. Pirard, J.-F. *et al.* HAWK-I: a new wide-field 1- to 2.5  $\mu$ m imager for the VLT. *Proc. SPIE* **5492**, 1763–1772 (2004).
58. Casali, M. *et al.* in *The New Era of Wide-Field Astronomy* (eds Clowes, R., Adamson, A. & Bromage, G.) 357–363 (ASPC Conf. Series, vol. 232, 2001).
59. Eastman, J., Siverd, R. & Gaudi, B. S. Achieving better than 1 minute accuracy in the heliocentric and barycentric Julian dates. *Publ. Astron. Soc. Pacif.* **122**, 935–946 (2010).
60. Mandel, K. & Agol, E. Analytic light curves for planetary transit searches. *Astrophys. J.* **580**, L171–L175 (2002).
61. Schwarz, G. Estimating the dimension of a model. *Ann. Stat.* **6**, 461–464 (1978).
62. Claret, A. & Bloemen, S. Gravity and limb-darkening coefficients for the Kepler, CoRoT, Spitzer, uvby, UVRIJK, and Sloan photometric systems. *Astron. Astrophys.* **529**, A75 (2011).
63. Gelman, A. & Rubin, D. B. Inference from iterative simulation using multiple sequences. *Stat. Sci.* **7**, 457–472 (1992).
64. Seager, S. & Mallén-Ornelas, G. A unique solution of planet and star parameters from an extrasolar planet transit light curve. *Astrophys. J.* **585**, 1038–1055 (2003).
65. Agol, E., Steffen, J., Sari, R. & Clarkson, W. On detecting terrestrial planets with timing of giant planet transits. *Mon. Not. R. Astron. Soc.* **359**, 567–579 (2005).
66. Davenport, J. R. A. *et al.* Kepler flares II: the temporal morphology of white-light flares on GJ 1243. *Astrophys. J.* **797**, 122 (2014).
67. Scargle, J. D. Studies in astronomical time series analysis. II. Statistical aspects of spectral analysis of unevenly spaced data. *Astrophys. J.* **263**, 835–853 (1982).
68. Goldreich, P. & Soter, S. Q in the solar system. *Icarus* **5**, 375–389 (1966).
69. Murray, C. D. & Dermott, S. F. *Solar System Dynamics* (Cambridge Univ. Press, 2001).
70. Limbach, M. A. & Turner, E. L. The orbital eccentricity—multiplicity relation and the solar system. *Proc. Natl Acad. Sci. USA* **112**, 20–24 (2015).
71. Chambers, J. E. A hybrid symplectic integrator that permits close encounters between massive bodies. *Mon. Not. R. Astron. Soc.* **304**, 793–799 (1999).
72. Charbonneau, D. *et al.* A super-Earth transiting a nearby low-mass star. *Nature* **462**, 891–894 (2009).
73. Miller-Ricci, E., Seager, S. & Sasselov, D. The atmospheric signatures of super-Earths: how to distinguish between hydrogen-rich and hydrogen-poor atmospheres. *Astrophys. J.* **690**, 1056–1067 (2009).
74. Han, E. *et al.* The exoplanet orbit database. II. Updates to exoplanet.org. *Publ. Astron. Soc. Pacif.* **126**, 827–837 (2014).
75. Kreidberg, L. *et al.* Clouds in the atmosphere of the super-Earth exoplanet GJ 1214b. *Nature* **505**, 69–72 (2014).
76. Berta, Z. K. *et al.* The flat transmission spectrum of the super-Earth GJ 1214b from Wide Field Camera 3 on the Hubble Space Telescope. *Astrophys. J.* **747**, 35 (2012).
77. Snellen, I. A. G., de Kock, R. J., de Mooij, E. J. W. & Albrecht, S. The orbital motion, absolute mass and high-altitude winds of exoplanet HD 209458b. *Nature* **465**, 1049–1051 (2010).
78. Rodler, F. & López-Morales, M. Feasibility studies for the detection of O<sub>2</sub> in an Earth-like exoplanet. *Astrophys. J.* **781**, 54 (2014).



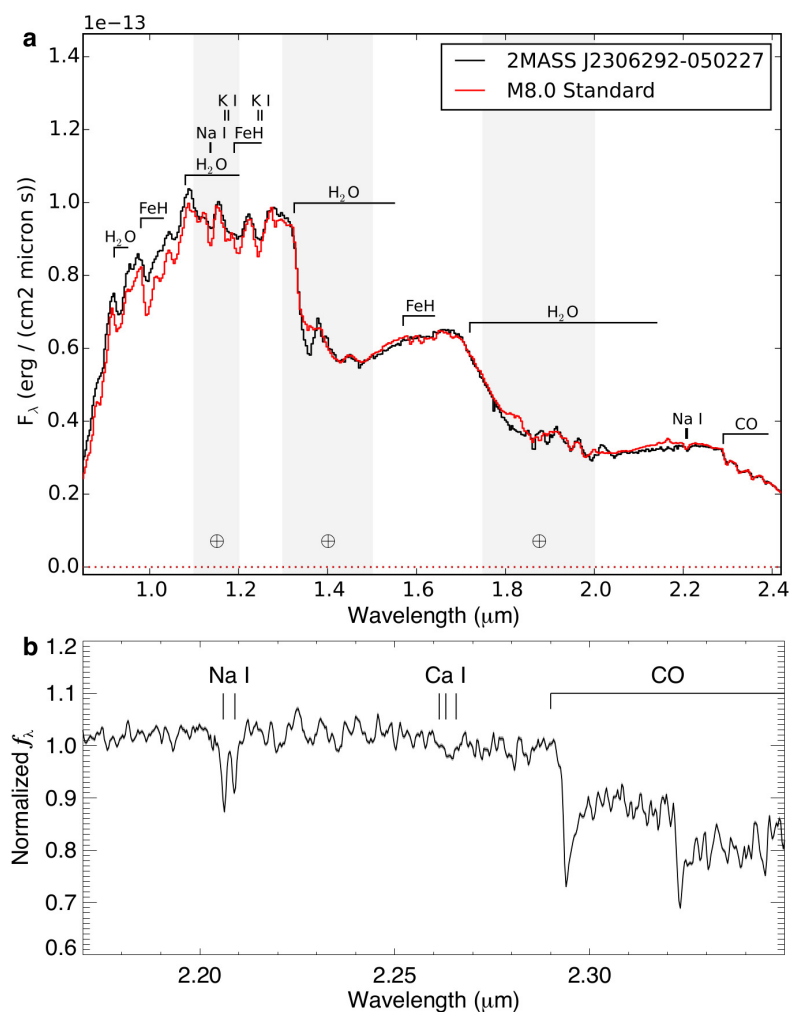
**Extended Data Figure 1 | Raw TRAPPIST-1 transit light curves.** The light curves are shown in chronological order from top to bottom and left to right, with unbinned data shown as cyan dots, and binned 0.005-day (7.2-minute) intervals shown as black dots with error bars. The error bars are the standard errors of the mean of the measurements in the bins.

The best-fit transit-plus-baseline models are overplotted (red line). The light curves are phased for the mid-transit time and shifted along the  $y$  axis for clarity. For the dual transit of 11 December 2015, the light curve is phased for the mid-transit time of planet TRAPPIST-1c. T1b, TRAPPIST-1b; T1c, TRAPPIST-1c; T1d, TRAPPIST-1d.



**Extended Data Figure 2 | De-trended TRAPPIST-1 transit light curves.** The details are as in Extended Data Fig. 1, except that the light curves shown here are divided by the best-fit baseline model to highlight the transit signatures.

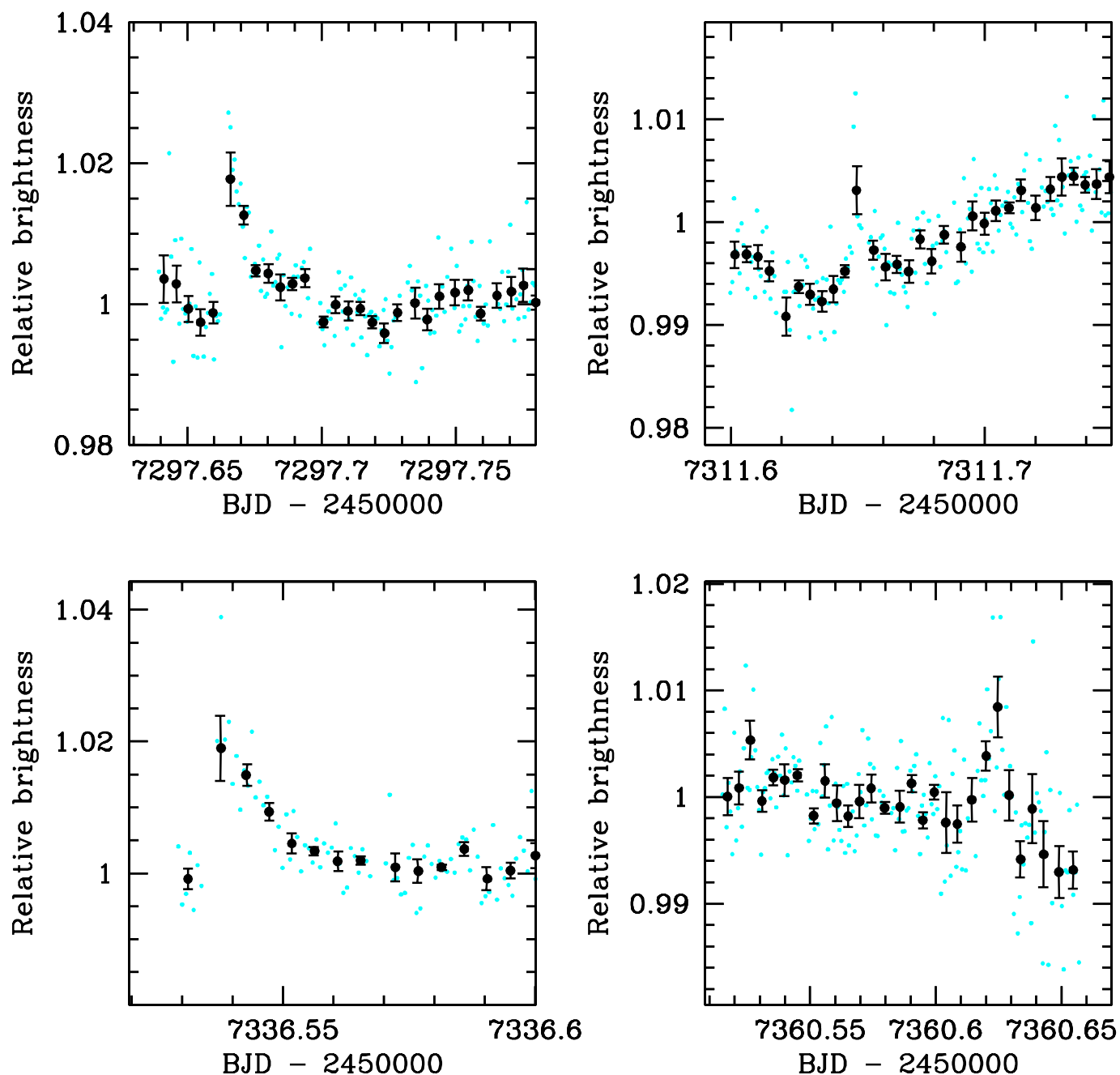




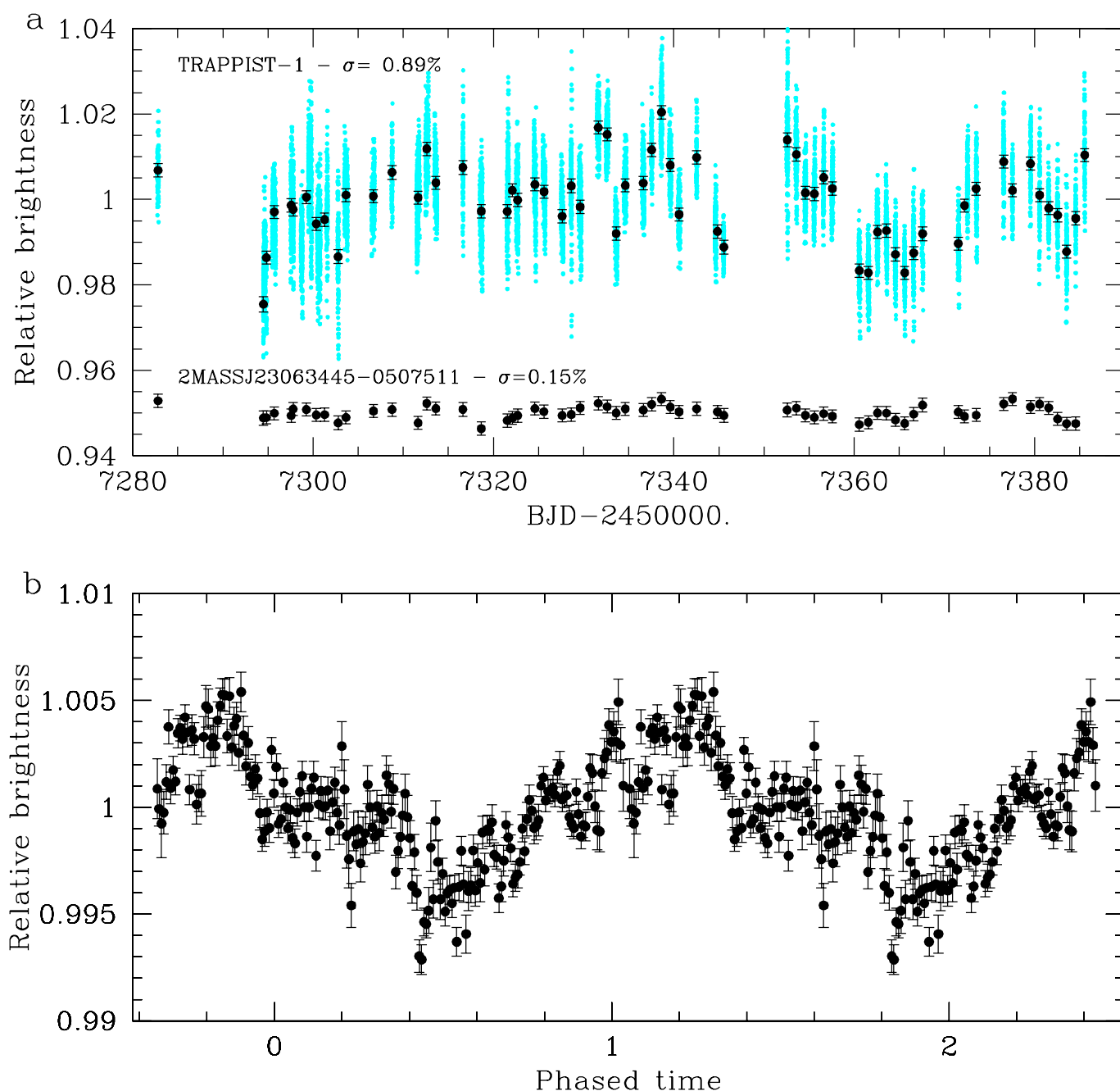
### Extended Data Figure 3 | Near-infrared spectra of TRAPPIST-1.

**a**, Comparison of TRAPPIST-1's near-infrared spectrum (black)—obtained with the spectrograph IRTF/Spex<sup>35</sup>—with that of the M8-type standard LHS132 (red). **b**, Cross-dispersed IRTF/Spex spectrum of

TRAPPIST-1 in the 2.17–2.35- $\mu\text{m}$  region. Na I, Ca I and CO features are labelled. Additional structure primarily originates from overlapping H<sub>2</sub>O bands. The spectrum is normalized at 2.2  $\mu\text{m}$ .  $F_{\lambda}$ , spectral flux density;  $f_{\lambda}$ , normalized spectra flux density.



Extended Data Figure 4 | Flare events in the TRAPPIST 2015 photometry. The photometric measurements are shown unbinned (cyan dots) and binned per 7.2-minute interval (black dots). For each interval, the error bars are the standard error of the mean.



#### Extended Data Figure 5 | Photometric variability of TRAPPIST-1.

**a**, Global light curve of the star as measured by TRAPPIST. The photometric measurements are shown unbinned (cyan dots) and binned per night (black dots with error bars ( $\pm$ s.e.m.)). This light curve is compared with that of the comparison star 2MASS J23063445 – 0507511,

shifted along the y axis for clarity. **b**, The same light curve for TRAPPIST-1, folded on the period  $P = 1.40$  days and binned by 10-minute intervals (error bars indicate  $\pm$ s.e.m.). For clarity, two consecutive periods are shown.



Extended Data Table 1 | TRAPPIST-1 transit light curves

Date	Instrument	Filter	$N_p$	$T_{\text{exp}}$	Baseline function	Transit(s)
17 Sep 2015	TRAPPIST	I+z	163	55s	$p(t^2+xy^2+f^2)$	TRAPPIST-1c
29 Sep 2015	TRAPPIST	I+z	232	55s	$p(t^2+xy^2+f^2)$	TRAPPIST-1d
27 Oct 2015	TRAPPIST	I+z	84	55s	$p(t^2+xy^2+f^2)$	TRAPPIST-1b
30 Oct 2015	TRAPPIST	I+z	77	55s	$p(t^2+xy^2+f^2)$	TRAPPIST-1b
05 Nov 2015	TRAPPIST	I+z	237	55s	$p(t^2+xy^2+f^2)$	TRAPPIST-1b
07 Nov 2015	TRAPPIST	I+z	241	55s	$p(t^2+xy^2+f^2)$	TRAPPIST-1c
08 Nov 2015	TRAPPIST	I+z	231	55s	$p(t^2+xy^2+f^2)$	TRAPPIST-1b
	VLT/HAWK-I	NB2090	207	17x1.7s	$p(t^2+xy^2+f^2)$	TRAPPIST-1b
11 Nov 2015	TRAPPIST	I+z	140	55s	$p(t^2+xy^2+f^2)$	TRAPPIST-1b
14 Nov 2015	TRAPPIST	I+z	241	55s	$p(t^2+xy^2+f^2)$	TRAPPIST-1b
18 Nov 2015	HCT/HFOSC	I	103	20s	$p(t^2+xy^2+f^2)$	TRAPPIST-1b
05 Dec 2015	UKIRT	J	1312	3x2s	$p(t^2+xy^2+f^2)$	TRAPPIST-1b
06 Dec 2015	UKIRT	J	1175	5x1s	$p(t^2+xy^2+f^2)$	TRAPPIST-1c
08 Dec 2015	UKIRT	J	1109	5x1s	$p(t^2+xy^2+f^2)$	TRAPPIST-1b
11 Dec 2015	TRAPPIST	I+z	158	55s	$p(t^2+xy^2+f^2)$	TRAPPIST-1c + d
28 Dec 2015	TRAPPIST	I+z	94	55s	$p(t^2+xy^2+f^2)$	TRAPPIST-1c (partial)

For each light curve, the date, instrument, filter, number of points ( $N_p$ ), exposure time ( $T_{\text{exp}}$ ), and baseline function are given. For the baseline functions,  $p(t^2)$ ,  $p(xy^2)$  and  $p(f^2)$  denote, respectively, second-order polynomial functions of time, of the x and y positions, and of the full-width at half-maximum of the stellar images.

Extended Data Table 2 | Quadratic limb-darkening coefficients

Bandpass	$u_1$	$u_2$
I (HCT/HFOSC)	$0.72 \pm 0.10$	$0.15 \pm 0.11$
I+z (TRAPPIST)	$0.65 \pm 0.10$	$0.28 \pm 0.12$
J (UKIRT/WFCAM)	$0.10 \pm 0.05$	$0.57 \pm 0.02$
NB2090 (VLT/HAWKI)	$0.04 \pm 0.03$	$0.50 \pm 0.03$

We inferred these values and errors for the quadratic coefficients  $u_1$  and  $u_2$  for TRAPPIST-1 from theoretical tables<sup>62</sup>, and used the values and errors as *a priori* knowledge of the stellar limb-darkening in a global MCMC analysis of the transit light curves. The error bars were obtained by propagation of the errors on the stellar gravity, metallicity, and effective temperature.

Extended Data Table 3 | Posterior likelihoods of the orbital solutions for TRAPPIST-1d

<b>TRAPPIST-1d period (d)</b>	<b>Circular orbit</b>	<b>Eccentric orbit</b>	<b><i>a</i> (au)</b>	<b><i>S<sub>p</sub></i> (<i>S<sub>Earth</sub></i>)</b>
4.551	0.0016	0.0017	0.023	0.98
5.200	0.0041	0.0045	0.025	0.82
8.090	0.012	0.013	0.034	0.45
9.101	0.018	0.011	0.037	0.39
10.401	0.139	0.0067	0.040	0.33
12.135	0.243	0.0029	0.045	0.26
14.561	0.393	0.0023	0.050	0.21
18.204	1	0.0018	0.058	0.15
24.270	0.212	0.0016	0.071	0.11
36.408	0.122	0.0014	0.093	0.06
72.820	$7.5e^{-5}$	$6.8e^{-8}$	0.147	0.02

The likelihoods shown for the circular and eccentric orbits are normalized to the most likely solution (that is, a circular orbit of  $P=18.204$  days (d)). For each orbit, the semi-major axis,  $a$  (in astronomical units (AU)), assuming a stellar mass of  $0.08M_{\odot}$  (Table 1), and the mean irradiation,  $S_p$  (in Earth units ( $S_{Earth}$ )) are shown.



Extended Data Table 4 | Individual mid-transit timings measured for the TRAPPIST-1 planets

Planet	Instrument	Epoch	Mid-transit timing (BJD <sub>TDB</sub> -2,450,000)
TRAPPIST-1b	TRAPPIST	0	$7322.5161^{+0.0013}_{-0.0010}$
	TRAPPIST	2	$7325.5391^{+0.0035}_{-0.0013}$
	TRAPPIST	6	$7331.5803 \pm 0.0013$
	TRAPPIST	8	$7334.6038 \pm 0.0012$
	VLT/HAWK-I	8	$7334.60490 \pm 0.00020$
	TRAPPIST	10	$7337.6249 \pm 0.0010$
	TRAPPIST	12	$7340.6474^{+0.0010}_{-0.0022}$
	HCT/HFOSC	15	$7345.18011 \pm 0.00089$
	UKIRT/WFCAM	26	$7361.79960 \pm 0.00030$
	UKIRT/WFCAM	28	$7364.82137 \pm 0.00056$
TRAPPIST-1c	TRAPPIST	0	$7282.8058 \pm 0.0010$
	TRAPPIST	21	$7333.6633 \pm 0.0010$
	UKIRT/WFCAM	33	$7362.72623 \pm 0.00040$
	TRAPPIST	35	$7367.5699 \pm 0.0012$
	TRAPPIST	42	$7384.5230 \pm 0.0011$
TRAPPIST-1d	TRAPPIST	0	$7294.7736 \pm 0.0014$
	TRAPPIST	?	$7367.5818 \pm 0.0015$

The transit timings shown were deduced from individual analyses of the transit light curves, assuming circular orbits for the planets. The error bars correspond to the  $1\sigma$  limits of the posterior PDFs of the transit timings.

# Lightwave-driven quasiparticle collisions on a subcycle timescale

F. Langer<sup>1</sup>, M. Hohenleutner<sup>1</sup>, C. P. Schmid<sup>1</sup>, C. Poellmann<sup>1</sup>, P. Nagler<sup>1</sup>, T. Korn<sup>1</sup>, C. Schüller<sup>1</sup>, M. S. Sherwin<sup>2</sup>, U. Huttner<sup>3</sup>, J. T. Steiner<sup>3</sup>, S. W. Koch<sup>3</sup>, M. Kira<sup>3</sup> & R. Huber<sup>1</sup>

Ever since Ernest Rutherford scattered  $\alpha$ -particles from gold foils<sup>1</sup>, collision experiments have revealed insights into atoms, nuclei and elementary particles<sup>2</sup>. In solids, many-body correlations lead to characteristic resonances<sup>3</sup>—called quasiparticles—such as excitons, dropletions<sup>4</sup>, polarons and Cooper pairs. The structure and dynamics of quasiparticles are important because they define macroscopic phenomena such as Mott insulating states, spontaneous spin- and charge-order, and high-temperature superconductivity<sup>5</sup>. However, the extremely short lifetimes of these entities<sup>6</sup> make practical implementations of a suitable collider challenging. Here we exploit lightwave-driven charge transport<sup>7–24</sup>, the foundation of attosecond science<sup>9–13</sup>, to explore ultrafast quasiparticle collisions directly in the time domain: a femtosecond optical pulse creates excitonic electron–hole pairs in the layered dichalcogenide tungsten diselenide while a strong terahertz field accelerates and collides the electrons with the holes. The underlying dynamics of the wave packets, including collision, pair annihilation, quantum interference and dephasing, are detected as light emission in high-order spectral sidebands<sup>17–19</sup> of the optical excitation. A full quantum theory explains our observations microscopically. This approach enables collision experiments with various complex quasiparticles and suggests a promising new way of generating sub-femtosecond pulses.

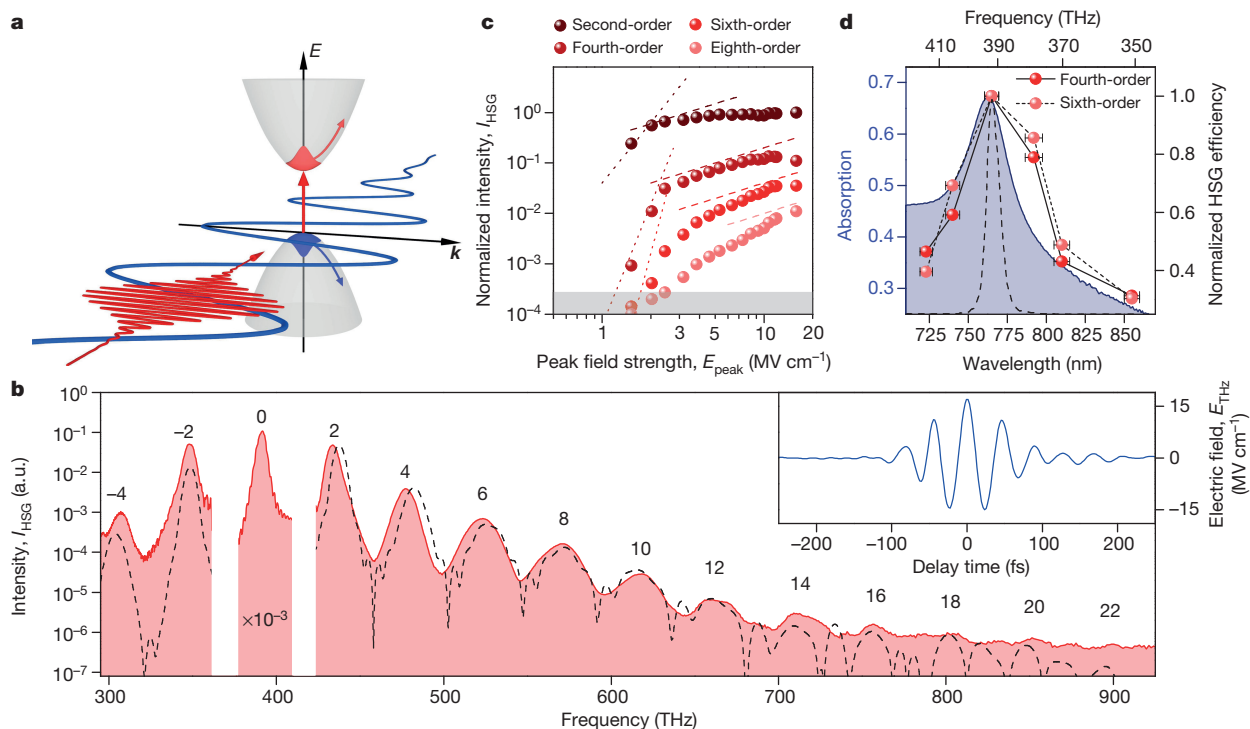
Following the principle of high-energy accelerators, their solid-state counterparts should prepare, accelerate and collide quasiparticles while detecting the outcomes. This can be realized only if all steps are accomplished faster than the ultrashort lifetime of the quasiparticle<sup>6</sup>. Preparing the quantum states of quasiparticles by femtosecond lasers is already well established<sup>25</sup>. Ultrafast acceleration, in turn, may harness the latest breakthroughs of strong-field control: the carrier wave of an intense laser pulse has been applied to ionize atoms or molecules and recollide their fragments when the sign of the light-field changes<sup>10</sup>. The excess energy of the collision is released as high-order harmonics of the laser frequency, which may emerge in attosecond bursts<sup>9</sup>, carrying key structural and dynamical information about the participants<sup>11,12</sup>. Recently, this idea of subcycle control has been extended to solids, in which lightwaves have driven dynamical Bloch oscillations<sup>20,21</sup>, interband quantum interference<sup>24</sup> and high-order harmonic generation<sup>20–24</sup> (HHG). However, the microscopic mechanisms are complex because the lightwave drives an inseparable combination of interband excitation<sup>21,24,26</sup> (quasiparticle preparation) and intraband acceleration<sup>20–24,26</sup>. The recent observation of high-order sideband generation (HSG) in gallium arsenide quantum wells<sup>18</sup> has suggested a way to disentangle these steps. Following optical preparation of coherent excitons<sup>27</sup>, a strong terahertz (THz) field of frequency  $\nu_{\text{THz}}$  modulates the interband resonance such that spectral sidebands at even multiples of  $\nu_{\text{THz}}$  emerge<sup>17–19</sup>. This observation has been modelled by THz-driven electron–hole collisions, similarly to the three-step model of HHG in atomic gases<sup>7,8</sup>. Yet, the potential of HSG for real-time quasiparticle collisions has remained untapped.

Here, we introduce a subcycle quasiparticle collider. A femtosecond optical pulse prepares coherent excitons in tungsten diselenide (WSe<sub>2</sub>), at a well-defined phase of a THz driving field ( $E_{\text{THz}}$ ). This phase defines how  $E_{\text{THz}}$  accelerates and collides electrons and holes, thereby yielding photon emission by pair annihilation, which is detected as high-order sidebands. We reveal the subcycle quantum dynamics of quasiparticle collisions, model the microscopic process by a full quantum theory and use the data to extract key information on the colliding species. Finally, we compare the process to HHG in the same material.

As a member of the class of transition metal dichalcogenides, WSe<sub>2</sub> features strongly bound Wannier excitons<sup>27,28</sup> and unique spin–valley coupling<sup>29</sup>. Even in its bulk form, WSe<sub>2</sub> has been reported to exhibit an exciton binding energy of the order of 0.1 eV at its direct bandgap<sup>30</sup>, which implies that these quasiparticles are stable at room temperature. In a first experiment, a 100-fs excitation pulse (red waveform in Fig. 1a) centred at a photon energy of 1.621 eV (centre frequency, 392 THz) resonantly prepares coherent excitons at the direct band gap of WSe<sub>2</sub> (thickness, 60 nm) located at the K and K' points in momentum space. A co-propagating intense THz pulse (blue waveform in Fig. 1a, b; centre photon energy,  $h\nu_{\text{THz}} = 95$  meV ( $h$  is the Planck constant); centre frequency,  $\nu_{\text{THz}} = 23$  THz; peak electric field in air,  $E_{\text{peak}} = 17$  MV cm<sup>−1</sup>; see Methods section ‘Experimental set-up’) modulates the excitonic polarization and gives rise to HSG<sup>18,19</sup>. The resulting intensity spectrum  $I_{\text{HSG}}$  includes sidebands of positive and negative orders  $n$ , spaced at  $2\nu_{\text{THz}}$  (Fig. 1b). Spectral components up to  $n = 22$  are clearly visible, while the intensities of negative orders drop much faster with decreasing frequencies. The new frequency components feature the same polarization as the interband excitation pulse (Extended Data Fig. 1). In contrast to perturbative scaling ( $I_{\text{HSG}} \propto E_{\text{peak}}^{2n}$ ), the sideband intensity grows linearly with  $E_{\text{peak}}$  for intermediate peak fields, depending on the sideband order, and even saturates at the highest field strengths (Fig. 1c, see also Extended Data Fig. 2 for the influence of the interband excitation fluence). Finally,  $I_{\text{HSG}}$  peaks if the interband excitation is resonant with the excitonic absorption peak (Fig. 1d), underpinning its excitonic origin. Apart from the broadening of the sidebands due to the ultrashort duration of  $E_{\text{THz}}$ , these findings agree well with previous time-integrated studies<sup>18</sup>.

To directly resolve the underlying subcycle quantum dynamics, we now prepare the coherent excitons within 10 fs—much faster than a single oscillation period  $T = \nu_{\text{THz}}^{-1}$  of the THz wave. Figure 2a depicts the spectral shape of the observed sideband intensity  $I_{\text{HSG}}$  as a function of the delay between the excitation pulse and the peak of the THz driving field,  $t_{\text{ex}}$ . Electro-optic detection allows us to determine the complete waveform  $E_{\text{THz}}$  on the same absolute timescale (Fig. 2a, b, blue curve). The spectrogram in Fig. 2a follows a nearly Gaussian temporal envelope, resembling the THz envelope. Interestingly,  $I_{\text{HSG}}$  is strongly modulated along the time axis with a period of  $T/2$ ; that is, there is a subcycle criterion for ‘good’ and ‘bad’ preparation times  $t_{\text{ex}}$ .

<sup>1</sup>Department of Physics, University of Regensburg, 93040 Regensburg, Germany. <sup>2</sup>Department of Physics and the Institute for Terahertz Science and Technology, University of California at Santa Barbara, Santa Barbara, California 93106, USA. <sup>3</sup>Department of Physics, University of Marburg, 35032 Marburg, Germany.



**Figure 1 | High-order sideband generation in tungsten diselenide.**

**a**, Schematic of the experiment in reciprocal space ( $E$ , energy;  $k$ , wave vector): an interband excitation pulse (red waveform) creates an excitonic polarization in WSe $_2$  (red, vertical arrow), while a strong multi-THz field (blue waveform) simultaneously accelerates the wave packets of the electron and hole (curved arrows) within their respective bands (grey shaded parabolas). **b**, Measured intensity spectrum (red) of high-order sidebands from WSe $_2$  (thickness, 60 nm; sample at room temperature) driven by a phase-locked THz transient featuring a centre frequency of 23 THz and an external peak field strength of  $E_{\text{THz}} = 17 \text{ MV cm}^{-1}$  (see inset; Keldysh parameter  $\gamma = 0.08 < 1$ ) under resonant optical excitation at a frequency of 392 THz (labelled by '0', 0th order has been multiplied by a

factor of  $10^{-3}$ ; other numerals denote the order of sidebands). The black dashed curve shows the calculated intensity spectrum  $I_{\text{HSG}}$ , a.u., arbitrary units. **c**, Recorded high-order sideband intensity  $I_{\text{HSG}}$  of orders ( $n$ ) two to eight as a function of driving peak field strength  $E_{\text{peak}}$ . Dotted lines follow a perturbative scaling law,  $I_{\text{HSG}} \propto E_{\text{peak}}^{2n}$ ; dashed lines mark a linear scaling; the grey shaded area indicates the noise level. **d**, Measured generation efficiency of the fourth- and sixth-order sideband for different excitation wavelengths (red spheres). Error bars represent the bandwidth (full-width at half maximum) of the excitation spectra (excitation spectrum at 392 THz shown as a dashed curve). The shaded area shows the measured exciton resonance in the absorption spectrum of the sample.

For a quantitative timing analysis, Fig. 2b compares  $E_{\text{THz}}$  (blue curve) with the spectrally integrated intensity  $I_{\text{HSG}}$  (red shaded curve). The strongest THz half-cycle occurs at a delay  $\delta_{\text{global}} \approx T$  with respect to the most-intense sideband peak. This retardation rules out instantaneous optical nonlinearities as the microscopic origin of HSG. A detailed comparison of  $E_{\text{THz}}$  and  $I_{\text{HSG}}$  also reveals a subcycle delay  $\delta_{\text{sc}}$  of the observed sideband bursts relative to the THz crests (see inset in Fig. 2b). For subsequent half-cycles of the driving field,  $\delta_{\text{sc}}$  initially increases, reaches a maximum value of approximately  $T/8$  at the centre of the THz pulse and slightly decreases afterwards (Fig. 2d).

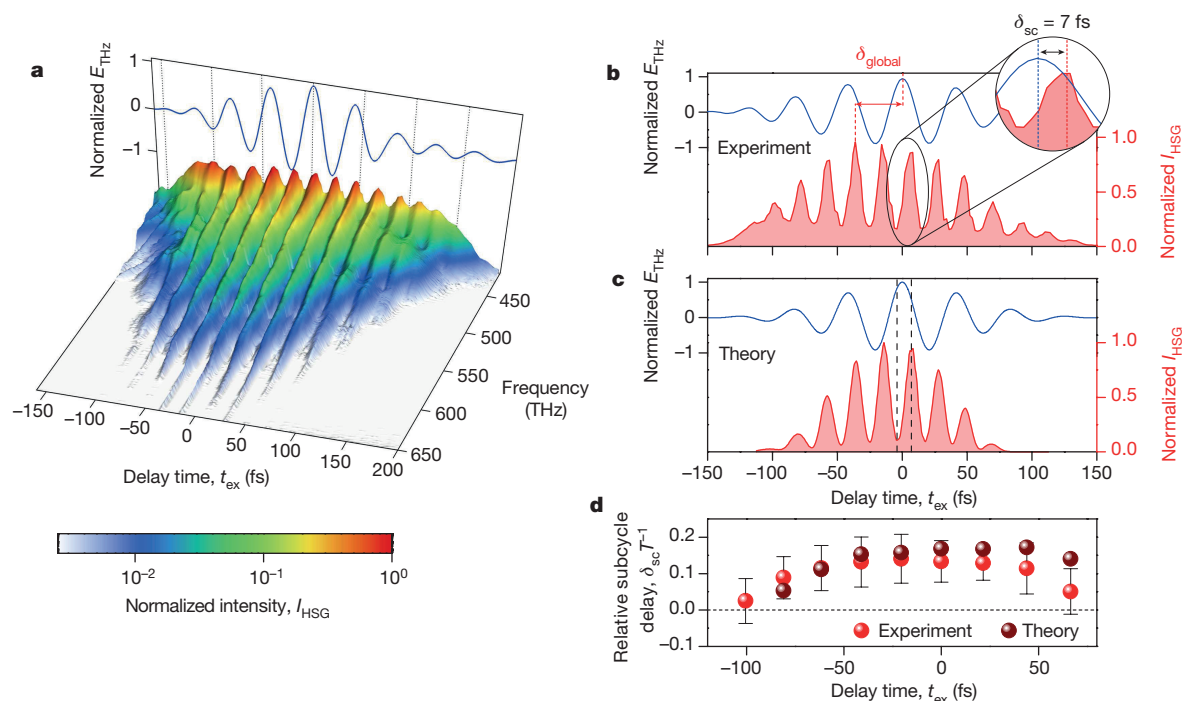
To link these signatures of  $I_{\text{HSG}}$  with the microscopic quasiparticle motion, we analyse the experiment using a full quantum mechanical model based on the semiconductor Bloch equations (see Methods section 'Microscopic model of HSG'). For a 100-fs preparation pulse, the experimentally observed HSG spectrum is well reproduced by our calculations (Fig. 1b, black dashed curve), showing a similar modulation and intensity roll-off on the high- and low-frequency sides. Using a 10-fs preparation pulse, the calculations quantitatively reproduce the temporal characteristics of  $I_{\text{HSG}}$  (Fig. 2c), including even delicate features such as  $\delta_{\text{sc}}$  (Fig. 2d).

Figure 3 visualizes the underlying microscopic quasiparticle dynamics for two characteristic preparation times  $t_{\text{ex}}$ , corresponding to minimum ( $t_{\text{ex}} = -4 \text{ fs}$ , Fig. 3c, e, g) and maximum ( $t_{\text{ex}} = 7 \text{ fs}$ , Fig. 3d, f, h) HSG. The two quasiparticle-collider sequences are schematically illustrated by Feynman diagrams in Fig. 3a, b. For  $t_{\text{ex}} = -4 \text{ fs}$ , the coherent excitons are prepared shortly before a THz field crest (see vertical dashed lines in Figs 2c and 3c). Immediately after optical preparation, the average momenta of the electron (Fig. 3e, solid curve)

and the hole (dashed curve) are located in the vicinity of the K and K' points of the Brillouin zone (defined as  $k = 0$ ). The strong THz field accelerates the distribution function of the electron ( $f_k^e$ ; Fig. 3e, false-colour plot) and the hole ( $f_k^h$ ; not shown) away from  $k = 0$ , driving the electron and hole wave packets apart. A qualitatively different situation occurs for  $t_{\text{ex}} = 7 \text{ fs}$  (Fig. 3d, f, h), for which the coherent excitons are prepared shortly before a zero crossing of the THz field (vertical dashed lines in Figs 2c and 3d). Now the mean momenta change sign; that is, the initially separating electron and hole are set on a recollision path (Fig. 3f).

Particle preparation of typical accelerators is spatially precise and collisions are monitored in real space. Our concept follows the relative motion of quasiparticles, and the preparation is precise in terms of the relative coordinates of the quasiparticles because the optical pulse creates them in a single, spatially correlated excitonic state described by a pair correlation function<sup>27</sup> (see Methods section 'Coherent electron-hole pair-correlation function'). Figure 3g, h displays its coherent part  $g_{\text{coh}}(r)$ , which defines how electrons are coherently coupled to holes as a function of their relative distance  $r$  (reciprocal space representation in Extended Data Fig. 3). In both panels,  $g_{\text{coh}}(r)$  is maximum close to the peak of the optical pulse before it decays by ultrafast dephasing (dephasing time  $T_2 < T/2$ ). Figure 3c, d shows the corresponding mean electron-hole separation ( $r$ ). Suitable  $t_{\text{ex}}$  (Fig. 3d) allows the THz field (Fig. 3c, d, blue curve) to separate and collide the electron-hole pairs at high energy. Owing to this ballistic motion, maximum HSG occurs for preparation times that are offset with respect to the field crests, causing the distinct delay  $\delta_{\text{sc}}$  observed in Fig. 2d. The precise temporal structure seen in the experiment can be reproduced only if





**Figure 2 | Subcycle electron–hole recollisions.** **a**, Spectrally resolved high-order sideband intensity  $I_{\text{HSG}}$  (colour scale) as a function of delay time  $t_{\text{ex}}$  between the THz driving field ( $E_{\text{THz}}$ , blue curve, see also **b**) and a 10-fs interband excitation pulse. **b**, **c**, Measured (**b**) and calculated (**c**) high-order sideband intensity  $I_{\text{HSG}}$  (red, spectrally integrated between 435 THz and 650 THz) on the same timescale as the driving waveform ( $E_{\text{THz}}$ , blue), which peaks with a global delay  $\delta_{\text{global}}$  after  $I_{\text{HSG}}$ . On a subcycle

scale, the recorded sideband intensity (red) peaks at a distinct time delay  $\delta_{\text{sc}}$  after the nearest extrema of the driving waveform (see close-up in **b**). **d**, Subcycle delay of  $I_{\text{HSG}}$  in units of the driving period  $T$  for subsequent driving half-cycles at their respective delay times as measured (bright red spheres; error bars, standard deviation of  $\delta_{\text{sc}}$  for 25 consecutive measurements) and calculated (dark red spheres). The horizontal black dashed line marks  $\delta_{\text{sc}} T^{-1} = 0$ .

excitation-induced dephasing<sup>27</sup> is taken into account (see Methods section ‘Effect of excitation-induced dephasing on HSG’ and Extended Data Figs 4–6). By contrast, for ‘bad’ timing  $t_{\text{ex}}$ , the electron and the hole are monotonically driven apart (Fig. 3c), suppressing collisions (see Fig. 3a). The situation is analogous to a cyclotron, in which electrons are effectively accelerated only if they are injected at the right phase of the alternating accelerator field.

Whereas the average intraband dynamics  $\langle r \rangle$  resembles semiclassical trajectories (see Fig. 3c, d), the connection between quasiparticle collision and HSG requires a full quantum theory. Similarly to electron–positron pairs in vacuum, colliding electron–hole pairs in a solid are subject to pair annihilation, during which a high-energy photon may be emitted, leading to sideband signals; this microscopic annihilation dynamics manifests itself in  $g_{\text{coh}}(r)$  (Fig. 3h). After substantial spreading of  $g_{\text{coh}}(r)$ , electron–hole collision is prompted by an abrupt termination of the correlation function due to pair annihilation. Subsequently, interference patterns occur among the residual fragments that survive the collision (see black ellipse in Fig. 3h). These distinct collision-induced annihilation features are absent in the case of ‘bad’ injection times, for which the coherence gradually abates owing to dephasing (Fig. 3g). The details of  $g_{\text{coh}}(r)$  are a direct manifestation of the many-body interactions governing the internal structure of the exciton, and are evident in the temporal and spectral response in HSG.

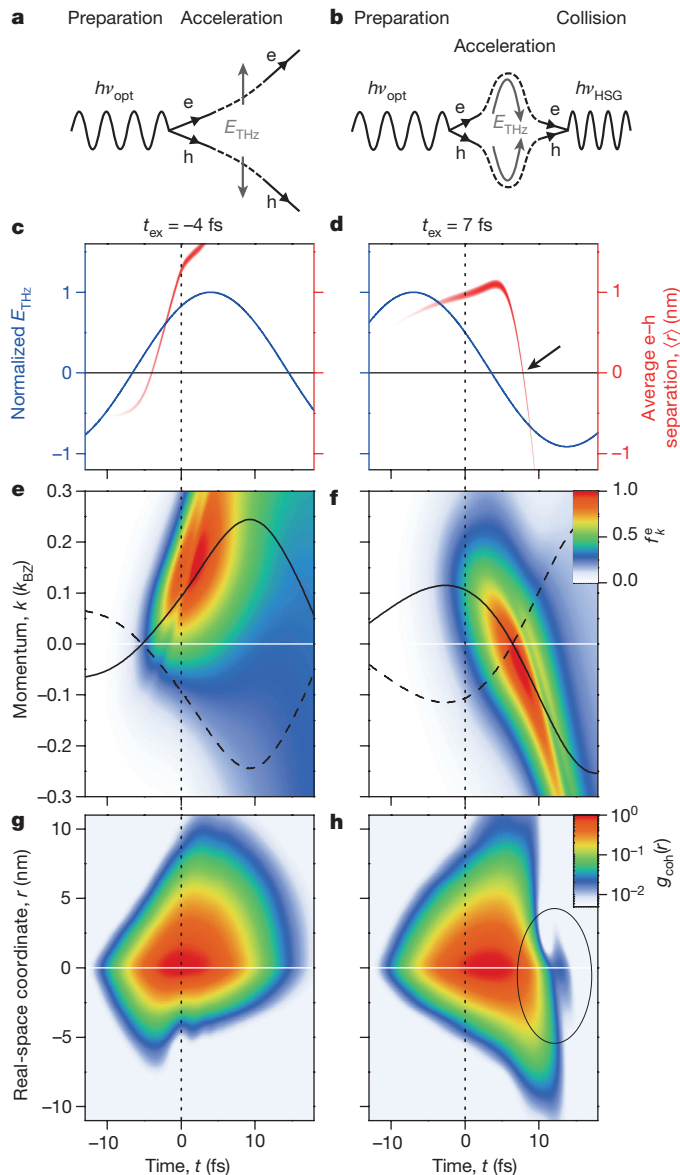
Detection of HSG, combined with precise preparation and acceleration, delivers an in-depth view of quasiparticles. As in particle colliders, one can expose the scattering products by systematic variation of key parameters, such as the strength, frequency and duration of the preparation and THz pulses, and their relative delay  $t_{\text{ex}}$ . In Methods section ‘Data landscape of quasiparticle colliders’, we provide first examples of the ways in which to retrieve material-specific information such as the exciton binding energy, scattering times and Coulomb enhancement of the collisional cross-section via dedicated analyses of the multi-dimensional data landscape of  $I_{\text{HSG}}$ . In particular, the collision events

recorded in our experiments bear the hallmarks of an exciton binding energy around 60 meV.

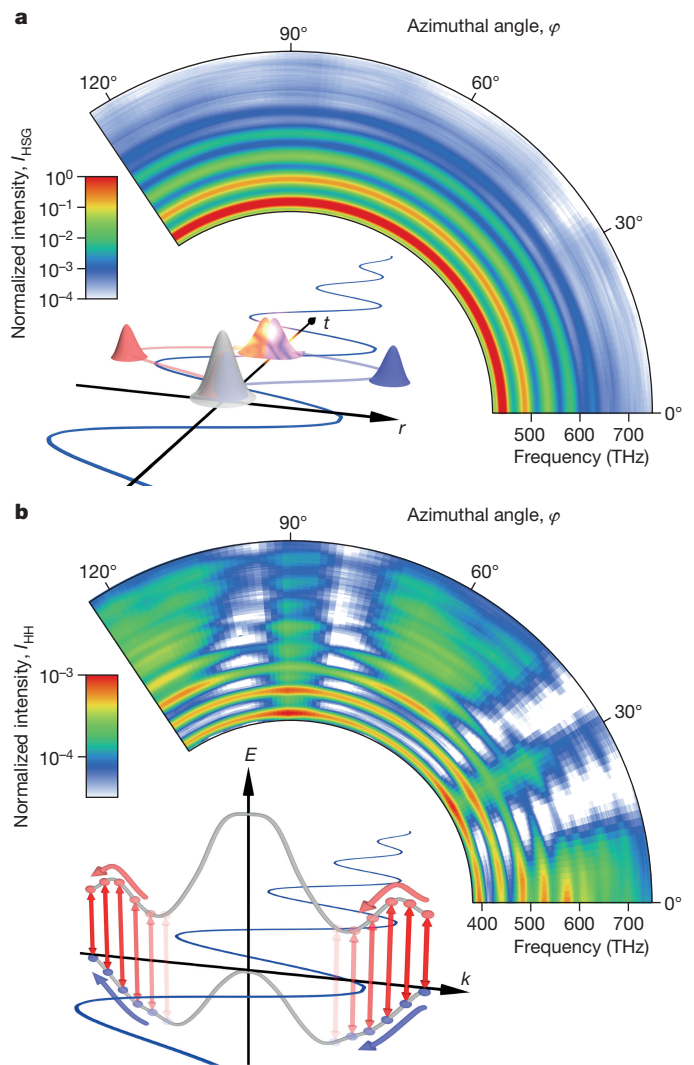
Well-defined initial momenta and relative spatial coordinates of the electrons and holes are prerequisites for controlled collisions. By contrast, the direction of acceleration in the WSe<sub>2</sub> crystal is less relevant, as seen in Fig. 4: the high-order sideband spectra show virtually no dependence on the azimuthal in-plane angle  $\varphi$  of the THz polarization. This behaviour is compatible with our theoretical expectation that coherent electron–hole wave packets in HSG remain rather close to  $k = 0$  (Fig. 3f). Both Coulomb attraction between electrons and holes and fast dephasing keep the coherent electron–hole excursions within a region in which the effective mass is approximately isotropic (see Fig. 1a). Additionally, HSG probes the interband polarization of colliding electrons and holes only at spatial overlap<sup>27</sup> (inset of Fig. 4a).

This unique characteristic of the quasiparticle collider becomes particularly clear when we compare the dynamics with HHG in the same sample, which is achieved by applying the strong THz field without optical preparation of excitons. A typical high-harmonic spectrum exhibiting odd orders up to  $n' = 47$  is depicted in Extended Data Fig. 7. The intensity of HHG is smaller by approximately one order of magnitude than HSG for frequencies below 650 THz. Owing to the non-perturbative, non-resonant excitation of Bloch electrons with simultaneous acceleration by the strong-field bias<sup>20,21,24,26</sup>, the initial carrier distribution is smeared out throughout the entire Brillouin zone. Furthermore, coherent interband polarization can emit photons at any point in momentum space (see inset of Fig. 4b). This situation manifests itself by a critical dependence of the intensity and the shape of high-order harmonics on  $\varphi$  (Fig. 4b), reflecting the six-fold symmetry of WSe<sub>2</sub> in the layer plane.

In conclusion, we demonstrate the realization of quasiparticle collisions in the time domain. We excite, accelerate and collide coherent electron–hole pairs in WSe<sub>2</sub> resulting in ultrabroad sideband emission. On a subcycle scale, we directly trace the retarded, ballistic wave-packet



**Figure 3 | Quantum simulation of subcycle electron-hole collisions underlying HSG.** **a, b**, Schematic Feynman diagrams depicting electron-hole (e-h) pair creation by a near-infrared photon ( $h\nu_{\text{opt}}$ ) and acceleration by  $E_{\text{THz}}$ : only for ‘good’ excitation times (**b**), the acceleration leads to a collision and annihilation of the electron-hole pair, thereby emitting a sideband photon  $h\nu_{\text{HSG}}$ . **c, d**, Trajectories (red, the intensity of the line represents the density of coherent excitons) tracing the real-time evolution (time  $t$ ) of the mean electron-hole separation (weighted average of  $g_{\text{coh}}(r)$ ) for characteristic delays  $t_{\text{ex}}$  corresponding to minimum (**c**) and maximum (**d**) HSG emission (compare with Fig. 2c). Vertical black dotted lines highlight  $t = 0$ , which marks the peak of the excitation pulse. Although electrons and holes are initially separated, they rapidly recollide (zero excursion marked by the black horizontal lines; time of recollision is highlighted by the black arrow in **d**) upon reversal of the driving field (blue curves), inducing a strong HSG signal (**d**). For ‘bad’ excitation times, the electron-hole separation increases monotonically, prohibiting recollisions (**c**). **e, f**, Occupation  $f_k^e$  (colour scale) of the first conduction band as a function of time  $t$  and crystal momentum  $k$  (in units of the wave vector  $k_{\text{BZ}}$  limiting the first Brillouin zone) for delays of  $t_{\text{ex}} = -4$  fs and 7 fs, respectively. Horizontal white lines mark  $k = 0$ . Black solid (dashed) curves trace the weighted average excursion of electrons (holes) in reciprocal space. **g, h**, Coherent electron-hole correlation function  $g_{\text{coh}}(r)$  (colour scale) as a function of time  $t$  and real-space coordinate  $r$ . Interference patterns occur after the abrupt collapse of coherence caused by electron-hole recollision (highlighted by the black ellipse in **h**), and are absent for  $t_{\text{ex}} = -4$  fs (**g**). Horizontal white lines mark  $r = 0$ .



**Figure 4 | Experimental comparison of high-order sideband and high-order harmonic generation.** **a, b**, Spectrally resolved high-order sideband intensity  $I_{\text{HSG}}$  (**a**; narrow-band exciton preparation at a frequency of 392 THz) and high-order harmonic intensity  $I_{\text{HH}}$  (**b**) as a function of the azimuthal orientation (angle  $\varphi$ ) of the  $\text{WSe}_2$  sample (normal incidence). Both intensity (colour) scales are normalized to the maximum HSG signal. Whereas high-order sidebands show virtually no dependence on  $\varphi$ , the high-harmonic spectra reflect the six-fold symmetry of  $\text{WSe}_2$ . **a**, Inset, real-space visualization of THz-driven electron-hole recollisions: upon excitation of a bound electron-hole pair (grey), the carriers (red and blue wave packets) are accelerated and recollided by the strong light field (blue) to recombine at  $r = 0$ , giving rise to high-order sideband emission. **b**, Inset, schematic of high-order harmonic generation: a polarization between valence and conduction band (grey) is induced by a strong multi-THz field (blue) and simultaneously accelerated within the bands (red and blue spheres, curved red and blue arrow). During this process, the coherent interband polarization is continuously modified (red vertical arrows).

dynamics and retrieve key information such as the exciton binding energy, scattering times and the Coulomb enhancement of the collisional cross-section. As in conventional particle accelerators, our optical excitation pulse prepares electron and hole wave packets with uncertainty-limited momenta and relative spatial coordinates. Yet their absolute position is defined only to within the relatively large focal spot size, which may be overcome in future near-field experiments by preparing excitons in a nanometre-sized interaction region. Both in the quasiparticle collider and in its large-scale counterpart, intrinsic characteristics of the collision become accessible by detecting the concomitantly emitted high-frequency radiation. The high

density of quasiparticles in solids helps to generate relatively intense, all-coherent, ultrabroadband radiation, which may pave the way to new sub-femtosecond light sources; furthermore, the subcycle modulation of HSG may facilitate information processing at optical-clock rates. Large-scale colliders have enabled the detailed study of elementary particles. Likewise, we anticipate that THz-driven, real-time collisions—operating on much smaller, widely tunable, length and energy scales—will be key to studying complex quasiparticles in numerous applications of modern materials science; examples include excitons, trions, biexcitons or dropletions in quasi-two-dimensional atomic monolayer systems, Dirac-like quasiparticles in graphene and topological insulators, or polarons in materials featuring strong electron–phonon coupling. This scheme may also be extended to optically excited quasiparticles in unconventional superconductors, and could help to resolve some of the outstanding questions in condensed matter research.

**Online Content** Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

**Received 13 November 2015; accepted 18 March 2016.**

- Rutherford, E. The scattering of  $\alpha$  and  $\beta$  particles by matter and the structure of the atom. *Philos. Mag. Ser. 6* **21**, 669–688 (1911).
- Wyatt, T. High-energy colliders and the rise of the standard model. *Nature* **448**, 274–280 (2007).
- Chemla, D. S. & Shah, J. Many-body and correlation effects in semiconductors. *Nature* **411**, 549–557 (2001).
- Almand-Hunter, A. E. *et al.* Quantum droplets of electrons and holes. *Nature* **506**, 471–475 (2014).
- Basov, D. N., Averitt, R. D., van der Marel, D., Dressel, M. & Haule, K. Electrodynamics of correlated electron materials. *Rev. Mod. Phys.* **83**, 471–541 (2011).
- Rossi, F. & Kuhn, T. Theory of ultrafast phenomena in photoexcited semiconductors. *Rev. Mod. Phys.* **74**, 895–950 (2002).
- Corkum, P. B. Plasma perspective on strong-field multiphoton ionization. *Phys. Rev. Lett.* **71**, 1994–1997 (1993).
- Lewenstein, M., Balcou, P., Ivanov, M. Y., L'Huillier, A. & Corkum, P. B. Theory of high-harmonic generation by low-frequency laser fields. *Phys. Rev. A* **49**, 2117–2132 (1994).
- Paul, P. M. *et al.* Observation of a train of attosecond pulses from high harmonic generation. *Science* **292**, 1689–1692 (2001).
- Corkum, P. B. & Krausz, F. Attosecond science. *Nature Phys.* **3**, 381–387 (2007).
- Li, W. *et al.* Time-resolved dynamics in  $\text{N}_2\text{O}_4$  probed using high harmonic generation. *Science* **322**, 1207–1211 (2008).
- Smirnova, O. *et al.* High harmonic interferometry of multi-electron dynamics in molecules. *Nature* **460**, 972–977 (2009).
- Neppl, S. *et al.* Direct observation of electron propagation and dielectric screening on the atomic length scale. *Nature* **517**, 342–346 (2015).
- Breuer, J. & Hommelhoff, P. Laser-based acceleration of nonrelativistic electrons at a dielectric structure. *Phys. Rev. Lett.* **111**, 134803 (2013).
- Wimmer, L. *et al.* Terahertz control of nanotip photoemission. *Nature Phys.* **10**, 432–436 (2014).
- Nanni, E. A. *et al.* Terahertz-driven linear electron acceleration. *Nature Commun.* **6**, 8486 (2015).
- Kono, J. *et al.* Resonant terahertz optical sideband generation from confined magnetoexcitons. *Phys. Rev. Lett.* **79**, 1758–1761 (1997).
- Zaks, B., Liu, R. B. & Sherwin, M. S. Experimental observation of electron–hole recollisions. *Nature* **483**, 580–583 (2012).
- Liu, R. B. & Zhu, B. F. High-order THz-sideband generation in semiconductors. *ALP Conf. Proc.* **893**, 1455–1456 (2007).
- Ghimire, S. *et al.* Observation of high-order harmonic generation in a bulk crystal. *Nature Phys.* **7**, 138–141 (2011).
- Schubert, O. *et al.* Sub-cycle control of terahertz high-harmonic generation by dynamical Bloch oscillations. *Nature Photon.* **8**, 119–123 (2014).
- Luu, T. T. *et al.* Extreme ultraviolet high-harmonic spectroscopy of solids. *Nature* **521**, 498–502 (2015).
- Vampa, G. *et al.* Linking high-harmonics from gases and solids. *Nature* **522**, 462–464 (2015).
- Hohenleutner, M. *et al.* Real-time observation of interfering crystal electrons in high-harmonic generation. *Nature* **523**, 572–575 (2015).
- Shah, J. *Ultrafast Spectroscopy of Semiconductors and Semiconductor Nanostructures* (Springer, Berlin 1999).
- Higuchi, T., Stockman, M. I. & Hommelhoff, P. Strong-field perspective on high-harmonic radiation from bulk solids. *Phys. Rev. Lett.* **113**, 213901 (2014).
- Kira, M. & Koch, S. W. *Semiconductor Quantum Optics* (Cambridge University Press, Cambridge, 2012).
- Poellmann, C. *et al.* Resonant internal quantum transitions and femtosecond radiative decay of excitons in monolayer  $\text{WSe}_2$ . *Nature Mater.* **14**, 889–893 (2015).
- Xu, X., Wang, Y., Xiao, D. & Heinz, T. F. Spin and pseudospins in layered transition metal dichalcogenides. *Nature Phys.* **10**, 343–350 (2014).
- Arora, A. Excitonic resonances in thin films of  $\text{WSe}_2$ : from monolayer to bulk material. *Nanoscale* **7**, 10421–10429 (2015).

**Acknowledgements** We thank H. Banks for critical reading of the manuscript. The work in Regensburg was supported by the European Research Council through grant number 305003 (QUANTUMsubCYCLE) and the Deutsche Forschungsgemeinschaft (through grant number HU 1598/2-1 and GRK 1570), and the work in Marburg by the Deutsche Forschungsgemeinschaft (through SFB 1083, SPP 1840 and grant numbers KI 917/2-2, KI 917/3-1). The work in Santa Barbara was supported by the National Science Foundation (through grant number DMR 1405964).

**Author Contributions** F.L., M.H., M.S.S., U.H., M.K. and R.H. conceived the study. F.L., M.H., C.P.S. and R.H. carried out the experiment and analysed the data. C.P., P.N., T.K. and C.S. provided, processed and characterized the samples. U.H., J.T.S., S.W.K. and M.K. developed the quantum-mechanical model and carried out the computations. All authors discussed the results and contributed to the writing of the manuscript.

**Author Information** Reprints and permissions information is available at [www.nature.com/reprints](http://www.nature.com/reprints). The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to R.H. ([rupert.huber@physik.uni-regensburg.de](mailto:rupert.huber@physik.uni-regensburg.de)) or M.K. ([mackillo.kira@physik.uni-marburg.de](mailto:mackillo.kira@physik.uni-marburg.de)).



## METHODS

**Experimental set-up.** Intense, phase-locked waveforms in the far- to mid-infrared spectral region (multi-terahertz range) are generated by difference frequency mixing of two spectrally detuned, near-infrared pulse trains from a dual optical parametric amplifier pumped by a titanium-sapphire laser amplifier<sup>21,24</sup>. The few-cycle transients feature peak fields of up to approximately  $1 \text{ V } \text{\AA}^{-1}$  (ref. 31). A super-continuum source based on filamentation in an yttrium aluminium garnet crystal delivers ultrabroadband white-light pulses covering the whole near-infrared and visible range. Both pulse trains are collinearly superimposed with an indium-tin-oxide-coated beam splitter and focused onto the tungsten diselenide sample under normal incidence using a gold-coated parabolic mirror. A mechanical delay stage in the excitation beam path allows us to temporally delay the pulses with respect to each other.

The sample under study is a 60-nm-thick sheet of  $\text{WSe}_2$  that has been cleaved using a viscoelastic gel film and placed on a diamond substrate grown by chemical vapour deposition. For resonant, narrow-band excitation of the excitonic polarization in  $\text{WSe}_2$ , optical band-pass filters featuring a transmittance bandwidth of 10 nm are used (see Fig. 1d). For the time-resolved measurements, the bandwidth of the excitation pulse is adjusted in the Fourier plane of a prism compressor, which enables the required pulse duration of 10 fs. Spectral components of the excitation pulse above 400 THz are filtered out to avoid the resonant generation of a high density of unbound electron-hole pairs. All experiments were performed at room temperature and in ambient air. No sample degradation was observed during the whole course of experiments.

The spot sizes of the excitation pulse and the THz pulse at the sample are  $22 \mu\text{m}$  and  $85 \mu\text{m}$ , respectively (intensity full-width at half-maximum). The generated high-order sideband intensity  $I_{\text{HSG}}$  and the high-order harmonic intensity  $I_{\text{HH}}$  are recorded with a spectrograph featuring a cooled silicon CCD camera. All spectra are corrected for the grating efficiency and the quantum efficiency of the detector.

The pulse energy of the narrowband excitation amounts to approximately 0.1 nJ in our current experiments. A comparison of the spectral components of the excitation and the sideband signals yields an energy conversion efficiency of approximately 0.1% in the 60-nm-thick  $\text{WSe}_2$  sample for the applied excitation fluence (approximately  $25 \mu\text{J cm}^{-2}$ ). Therefore, we estimate the energy in the high-order sideband pulse train as roughly 0.1 pJ.

Replacing the  $\text{WSe}_2$  sample by a 6.5- $\mu\text{m}$ -thick zinc telluride ( $\text{ZnTe}$ ) detector crystal allows for electro-optic detection of the driving field. To this end, we carefully place the crystal into the THz focus at the same position and account for the complex detector response<sup>32</sup>. The delay time  $t_{\text{ex}}$  between the THz waveform and the near-infrared excitation pulse enables a direct temporal correlation of the driving waveform with the recorded trace of high-order sideband emission. To describe the microscopic dynamics following exciton creation, a second timescale  $t$  is introduced, which describes the real time evolving for a fixed value of  $t_{\text{ex}}$  as shown, for instance, in Fig. 3.

**Microscopic model of HSG.** A complete quantum description of optical<sup>33,34</sup> and THz excitations<sup>21,24,35,36</sup> of solids follows from the semiconductor Bloch equations<sup>27</sup> (SBEs). Compared to earlier HSG models<sup>19</sup>, the SBEs fully include many-body effects beyond isolated electron-hole pairs, the interplay of HSG versus HHG, non-parabolic dispersions, and dephasing and relaxation processes. The SBEs are

$$\begin{aligned} i\hbar \frac{\partial}{\partial t} p_k &= \tilde{\varepsilon}_k p_k - \hbar \Omega_k(t) (1 - f_k^e - f_k^h) + i|e|E_{\text{THz}}(t) \cdot \nabla_k p_k + I_k^c \\ \hbar \frac{\partial}{\partial t} f_k^e &= -2\text{Im}[\hbar \Omega_k(t) p_k^*] + |e|E_{\text{THz}}(t) \cdot \nabla_k f_k^e + I_k^c \\ \hbar \frac{\partial}{\partial t} f_k^h &= -2\text{Im}[\hbar \Omega_{-k}(t) p_{-k}^*] + |e|E_{\text{THz}}(t) \cdot \nabla_{-k} f_{-k}^h + I_k^h \end{aligned} \quad (1)$$

in which  $p_k^*$  is the complex conjugate of the microscopic interband polarization  $p_k$  and  $f_k^e$  ( $f_k^h$ ) defines the electron (hole) occupation in a conduction-band (valence-band) state with a crystal momentum  $\hbar k$  ( $\hbar$  is the reduced Planck constant). The equations also contain the strength of the field coupling via  $\Omega_k$  and the renormalized single-particle energies  $\tilde{\varepsilon}_k$ ; the elementary charge is denoted by  $e$ , and scattering contributions are considered by  $I_k^c$ ,  $I_k^e$ , and  $I_k^h$ . Physically,  $p_k$  defines the transition amplitude of exciting an electron from one band to another while  $\hbar k$  is conserved. As an optical field  $E_{\text{opt}}(t)$  induces such a transition, it eventually generates occupations  $f_k^e$  and  $f_k^h$ . The presented SBEs can be generalized in a straightforward manner for multiple bands<sup>21,37,38</sup>.

As an inverse process, electrons recombine with holes via photon emission. The coherent part of this process yields radiative decay of the polarization<sup>27</sup> and the corresponding emission spectrum is<sup>35</sup>  $I = |\omega P(\omega) + iJ(\omega)|^2$  at energy  $\hbar\omega$ . The Fourier transforms of the macroscopic polarization  $P(t) = \sum_k d_{\text{cv}}(\mathbf{k}) p_k + \text{c.c.}$

(in which c.c. represents the complex conjugate) and the macroscopic current

$J(t) = \sum_{\lambda, \mathbf{k}} j_{\mathbf{k}}^{\lambda} f_{\mathbf{k}}^{\lambda}$  contain the dipole and current matrix elements  $d_{\text{cv}}(\mathbf{k})$  and  $j_{\mathbf{k}}^{\lambda} = \frac{ie}{\hbar} \nabla_{\mathbf{k}} \varepsilon_{\mathbf{k}}^{\lambda}$ , respectively. The currents in a given band  $\lambda = e, h$  are defined by its energy dispersion  $\varepsilon_{\mathbf{k}}^{\lambda}$ .

Because electrons and holes are charged particles, they experience a strong Coulomb force, which creates non-trivial many-body effects among  $p_k$ ,  $f_k^e$  and  $f_k^h$  during their finite lifetimes. For example, the strength of the field excitation becomes renormalized

$$\hbar \Omega_k(t) = d_{\text{cv}}(\mathbf{k}) \cdot \mathbf{E}(t) + \sum_{\mathbf{k}'} V_{\mathbf{k}-\mathbf{k}'} p_{\mathbf{k}'}$$

by the Coulomb matrix element  $V_{\mathbf{k}}$ , while its un-normalized part is a product of  $d_{\text{cv}}(\mathbf{k})$  and the total field  $\mathbf{E} = \mathbf{E}_{\text{opt}} + \mathbf{E}_{\text{THz}}$  containing also the THz field  $\mathbf{E}_{\text{THz}}$ . The Coulomb sum within  $\Omega_k(t)$  results in excitonic resonances in the polarization<sup>27</sup>. The single-particle transition energy  $\varepsilon_k$  becomes renormalized too

$$\tilde{\varepsilon}_k = \varepsilon_k - \sum_{\mathbf{k}'} V_{\mathbf{k}-\mathbf{k}'} (f_{\mathbf{k}'}^e + f_{\mathbf{k}'}^h)$$

Additionally, phonon, photon and Coulomb interactions yield scattering contributions  $\Gamma_k^e$ ,  $\Gamma_k^c$  and  $\Gamma_k^h$  due to high-order correlations<sup>33</sup>. We include microscopic scattering with the approximation  $\Gamma_k = -i \frac{\hbar}{T_2(\mathbf{k})} p_{\mathbf{k}}$ , which includes excitation-induced dephasing (EID) via the  $\mathbf{k}$ -dependent scattering time  $T_2(\mathbf{k})$ , as explained in Methods section 'Effect of excitation-induced dephasing on HSG'. Microscopically,  $\Gamma_k^e$  and  $\Gamma_k^h$  induce a relaxation of electron and hole distributions, which we implement using a phenomenological relaxation towards a symmetric carrier distribution via  $\Gamma_k^{\lambda} = -\frac{1}{2\tau} (f_k^{\lambda} + f_{-k}^{\lambda})$  on a timescale defined by  $\tau$  (ref. 24). In our computations,  $\tau$  is chosen to be 6 fs.

Our experiments show that high-order sideband generation (HSG) excites carriers only in the vicinity of the K point where  $\text{WSe}_2$  electrons and holes can be described with an effective mass approximation. To determine the general aspects of the interplay between HSG and high-harmonic generation (HHG), we use a one-dimensional (1D) two-band model<sup>35</sup>, and adjust  $V_k$  to produce a binding energy  $E_B = 60 \text{ meV}$  (corresponding to 14.5 THz) for the 1s-exciton state<sup>30</sup>. We match electrons and holes to produce the same effective mass  $m_e = m_h = 0.52 m_0$  (ref. 39), with the free electron mass  $m_0$ , close to the K point. We solve the closed system of coupled differential equations (1) numerically using a Runge-Kutta algorithm.

We have confirmed that a two-dimensional (2D) computation produces essentially the same HSG results. Note that bulk  $\text{WSe}_2$  features a layered structure where electron-hole motion is free only within each 2D layer. However, only a 1D computation is numerically feasible to fully include HHG, Bloch oscillations and a non-parabolicity in the dispersion  $\varepsilon_k$ . The 1D computations also confirm that HHG is an order of magnitude weaker than HSG, as in the experiment.

**Coherent electron-hole pair-correlation function.** The excitations underlying HSG can be investigated in real space by following the electron-hole pair-correlation function

$$g_{\text{eh}}(\mathbf{r}) \equiv \langle \hat{\rho}_e(\mathbf{r}) \hat{\rho}_h(\mathbf{0}) \rangle$$

containing the operators for electron and hole densities  $\hat{\rho}_e$  and  $\hat{\rho}_h$ , respectively. As discussed in ref. 33, the envelope of  $\hat{\rho}_e(\mathbf{r})$  and  $\hat{\rho}_h(\mathbf{0})$  determines the ways in which electrons move with respect to holes on length scales larger than the unit cell. The resulting  $g_{\text{eh}}(\mathbf{r})$  defines the conditional probability of finding an electron at position  $\mathbf{r}$  when a hole is at the origin.

To identify the coherent electron-hole acceleration, it is convenient to separate incoherent and coherent contributions from  $g_{\text{eh}}(\mathbf{r})$ . Following the derivation in ref. 33, we find  $g_{\text{eh}}(\mathbf{r}) = g_{\text{inc}}(\mathbf{r}) + g_{\text{coh}}(\mathbf{r})$ , with coherent part  $g_{\text{coh}}(\mathbf{r}) = |p(\mathbf{r})|^2$  and incoherent part  $g_{\text{inc}}(\mathbf{r}) = \rho_e \rho_h + \Delta g_{\text{eh}}(\mathbf{r})$ , which is proportional to the product of electron and hole densities ( $\rho_e$  and  $\rho_h$ ), to the lowest order. The formation of pairwise, incoherent correlations generates an additional contribution  $\Delta g_{\text{eh}}(\mathbf{r})$ . However,  $g_{\text{coh}}(\mathbf{r})$  is independent of correlation formation and is exclusively defined by the coherent polarization  $p(\mathbf{r})$ , which is given by the Fourier transform of  $p_k$  and describes the relative motion of electron-hole coherences. In other words,  $g_{\text{coh}}(\mathbf{r})$  defines the way in which coherently generated electrons are distributed with respect to the holes in real space. Because the SBEs fully include the many-body aspects of the polarization dynamics  $p_k$ ,  $g_{\text{coh}}(\mathbf{r})$  generalizes the descriptions of following a single electron-hole pair during HSG<sup>19,40</sup>.

**Effect of excitation-induced dephasing on HSG.** To quantify collision dynamics under different conditions, Extended Data Fig. 6b compares the mean wave vector

$$\langle \mathbf{k} \rangle = \frac{\sum_{\mathbf{k}} \mathbf{k} |p_{\mathbf{k}}|^2}{\sum_{\mathbf{k}} |p_{\mathbf{k}}|^2}$$

of the electron–hole wave packet for a 10-fs-long (red curve) and a 100-fs-long (black curve) optical excitation pulse centred at  $t_{\text{ex}} = 7$  fs, while the dephasing is kept constant ( $T_2 = 12$  fs). The THz waveform (Extended Data Fig. 6a) is identical for both optical excitations. The respective envelope of the excitation pulse is shown as the grey (yellow) shaded area for the 100-fs (10-fs) pulse.

We observe that the electron–hole wave packets are accelerated to much higher momentum states in the case of the long (100-fs) excitation pulse compared to the short (10-fs) excitation pulse. As discussed in the main text, a zero crossing of  $\langle k \rangle$  enables collisions and annihilation of optically generated coherent excitons. For the short excitation pulse, the magnitude of  $|p_k|^2$  substantially drops after the recollision. Therefore, the motion of residual coherent electron–hole pairs is indicated by the red dashed curve in Extended Data Fig. 6b, c, whereas the dominant  $|p_k|^2$  motion is plotted as red solid curve. For the long excitation pulse, the excitonic polarization  $|p_k|^2$  is generated continuously, such that it is not diminished substantially by the collisions. The corresponding mean wave vector  $\langle k \rangle$  is plotted as a black curve in Extended Data Fig. 6b, c.

Applying a constant dephasing time  $T_2$  implies that all excitonic components of the polarization decay with an identical rate. However, the Coulomb interaction scatters polarization with the excited electron–hole distributions, yielding a strong exciton-state-dependent EID<sup>33</sup>. More specifically, the 1s-exciton state (which is tightly bound around the K point) features lower scattering rates than other exciton states that spread out in  $k$ , which is included by introducing a  $k$ -dependent dephasing time  $T_2(k)$  as shown in the inset of Extended Data Fig. 4a: for low  $k$ ,  $T_2(k)$  is 12 fs, whereas it decreases to 2 fs at elevated  $k$ . Comparably fast dephasing times have also been observed in previous studies<sup>24,41,42</sup>. Additionally, the large temporal modulations in the HSG intensity as shown in Fig. 2a imply dephasing times  $T_2$  that are substantially shorter than one half-cycle of the THz driving field ( $T_2 \ll T/2 = 22$  fs).

To determine the effect of EID on electron–hole collisions, Extended Data Fig. 6c shows  $\langle k \rangle$  for the momentum-dependent dephasing  $T_2(k)$ . By comparing the black curves (100-fs-long excitation) in Extended Data Fig. 6b and c, we conclude that EID reduces the maximum excursion by roughly a factor of 3 compared to the constant dephasing result. The mean momentum trajectory  $\langle k \rangle$  for the short (10-fs) excitation pulse is less affected by EID, owing to the reduced maximum displacement.

Extended Data Fig. 4 compares the measured, time-integrated HSG spectra (shaded area) with computations including EID (black curves), a constant dephasing  $T_2 = 3.2$  fs (Extended Data Fig. 4a, red curve), and  $T_2 = 4$  fs (Extended Data Fig. 4b, blue curve), respectively. We observe that only calculations taking EID into account explain the height of all observed sidebands. More specifically, the simulations using a constant dephasing achieve good agreement only for either low-order sidebands ( $n < 8$ ,  $T_2 = 4$  fs) or sidebands above 600 THz ( $T_2 = 3.2$  fs). In numbers,  $T_2 = 4$  fs ( $T_2 = 3.2$  fs) overestimates (underestimates), for example, the tenth-order (fourth-order) sideband by a factor of 1.9 (1.7). This result implies that EID induces non-trivial spectral imprints on the HSG spectrum. These modulations result from the anharmonic electron–hole acceleration identified in Extended Data Fig. 6c.

In general, the SBE approach systematically describes the creation of coherent excitons, quasiparticle acceleration, collision-related annihilation, the interplay between HSG and HHG contributions, and the influence of many-body effects such as EID or phonon scattering processes as suggested by ref. 43. Only some aspects of these effects can be quantitatively analysed using either a semiclassical approach or a quantum description of a single electron–hole pair. Because many-body effects can induce drastic changes to HSG and HHG, it is clear that the SBE analysis is needed to properly parameterize any simplified model.

**Influence of excitation fluence.** With increasing optical fluence  $\Phi$ , the measured sideband intensity increases linearly (Extended Data Fig. 2a); however, the modulation depth of time-resolved traces of  $I_{\text{HSG}}$  is found to decrease (Extended Data Fig. 2b). This feature may be most intuitively explained by considering that a growing density of excitons may instantly lead to phase-space filling and broadening causing less-well-defined initial conditions for electron–hole recollisions. By contrast, a potential increase in dephasing rates does not seem to be a dominant factor, because faster dephasing should result in an increase in the modulation depth—the opposite effect of what we observe. Additionally, the subcycle delay  $\delta_{\text{sc}}$ , which is sensitively dependent on the dephasing time  $T_2$ , does not change substantially with varying excitation fluence, implying that major density-driven effects on  $T_2$  have not yet built up (Extended Data Fig. 2c). However, such excitation-induced effects are important for the case of narrowband excitation analysed in Fig. 1 and Extended Data Fig. 4.

**Data landscape of quasiparticle colliders.** In regular optical spectroscopy, system resonances of interest are often imbedded within the spectral range of the applied pump field. Our quasiparticle collider spectrally separates the relevant features from the pump as distinct sidebands. Additionally, each sideband order

can contain different aspects of the collisions. As in conventional colliders, the scattering-generated radiation produces a wealth of information about the (quasi) particles. Here, we study the HSG intensity  $I_{\text{HSG}}$  as a function of the HSG frequency  $\nu_{\text{HSG}}$ , the field strength  $E_{\lambda}$ , the duration  $\tau_{\lambda}$  and the central frequency  $\nu_{\lambda}$  of the optical preparation pulse ( $\lambda = \text{opt}$ ) and the THz waveform ( $\lambda = \text{THz}$ ), and their mutual delay  $t_{\text{ex}}$ , generating at least eight-dimensional datasets:  $I_{\text{HSG}}(\nu_{\text{HSG}}, E_{\text{THz}}, \tau_{\text{THz}}, \nu_{\text{THz}}, E_{\text{opt}}, \tau_{\text{opt}}, \nu_{\text{opt}}, t_{\text{ex}})$ . As is typical for colliders, such a wealth of data reveals details of colliding quasiparticles only through a dedicated identification process. For example, Fig. 3 demonstrates how changing  $t_{\text{ex}}$  yields a direct detection of the electron–hole recollision. We will next illustrate a few additional possibilities to extract exciton-related details from the HSG scattering data.

As shown in Extended Data Fig. 4, broadening of sideband resonances is directly related to exciton decay rates, which depend on the excitation parameters. This effect is often referred to as EID<sup>44,45</sup>, which can be controlled, for example, by changing  $E_{\text{opt}}$  (ref. 34). We have also confirmed that the exciton binding energy  $E_{\text{B}}$  strongly influences the overall strength of the HSG signal. For example, a tenfold increase of  $E_{\text{B}}$  amplifies  $I_{\text{HSG}}$  by roughly the same amount as a result of Coulomb enhancement of the collisional cross-section. Hence, a variety of material- and excitation-dependent exciton features may be deduced by quantitatively analysing the strength and width of  $I_{\text{HSG}}$  resonances.

By changing the optical pump frequency  $\nu_{\text{opt}}$  one can deduce differential changes in HSG spectra that sensitively monitor material details. We propose measuring a 2D  $I_{\text{HSG}}(\nu_{\text{HSG}}, \nu_{\text{opt}})$  set to follow spectral changes at the second-order sideband and to define a normalized spectrum

$$I(\nu, \nu_{\text{opt}}) \equiv \frac{I_{\text{HSG}}(\nu_{\text{HSG}} - \nu_{\text{opt}}, \nu_{\text{opt}})}{I_{\text{2SB}}}$$

in which  $\nu = \nu_{\text{HSG}} - \nu_{\text{opt}}$  is the frequency scale centred around  $\nu_{\text{opt}}$  and  $I_{\text{2SB}} = I_{\text{HSG}}(\nu_{\text{opt}} + 2\nu_{\text{THz}}, \nu_{\text{opt}})$  is the peak intensity of the second-order sideband. This normalization produces spectra that do not depend on the absolute intensity scale—a quantity that is difficult to determine experimentally for a wide range of frequencies  $\nu_{\text{opt}}$ . From the difference of two spectra measured with pump frequencies  $\nu_{\text{opt}} \pm \Delta\nu_{\text{opt}}/2$ , we deduce a finite differential

$$\Delta I(\nu, \nu_{\text{opt}}) = \frac{1}{\Delta\nu_{\text{opt}}} \left[ I(\nu, \nu_{\text{opt}} - \Delta\nu_{\text{opt}}/2) - I(\nu, \nu_{\text{opt}} + \Delta\nu_{\text{opt}}/2) \right]$$

which approaches the derivative  $-\frac{\partial I}{\partial \nu_{\text{opt}}}$  in the limit  $\Delta\nu_{\text{opt}} \rightarrow 0$ .

Extended Data Fig. 8a, b shows computed  $\Delta I(\nu, \nu_{\text{opt}})$  for two 1s-exciton binding energies,  $E_{\text{B}} = 60$  meV and  $E_{\text{B}} = 600$  meV, respectively, when  $h\nu_{\text{opt}}$  is expressed in terms of the detuning  $\Delta_{\text{opt}}$  with respect to the 1s-exciton energy  $h\nu_{1s}$  as  $h\nu_{\text{opt}} = h\nu_{1s} + \Delta_{\text{opt}}$ . These contours produce spectral information that differs greatly for the two exciton binding energies, demonstrating that  $\Delta I(\nu, \nu_{\text{opt}})$  exposes material-dependent properties. For  $E_{\text{B}} = 60$  meV,  $\Delta I(\nu, \nu_{\text{opt}})$  exhibits dominantly negative changes for  $\Delta_{\text{opt}} < 0$ , implying that  $I_{\text{HSG}}$  decreases for negative detuning. Because  $\Delta_{\text{opt}} > 0$  yields mainly positive  $\Delta I(\nu, \nu_{\text{opt}})$ , a clear dispersive shape is observed around the resonant excitation ( $\Delta_{\text{opt}} = 0$ ). For  $E_{\text{B}} = 600$  meV, a positive island appears within the  $\Delta_{\text{opt}} < 0$  region, which implies that  $I_{\text{HSG}}$  can be enhanced also for negative  $\Delta_{\text{opt}}$ , in contrast to the  $E_{\text{B}} = 60$  meV case. Additionally,  $\Delta_{\text{opt}} > 0$  produces a new feature because the positive islands split into a double-peaked structure for  $E_{\text{B}} = 600$  meV.

To analyse the quantitative dependence of  $\Delta I(\nu, \nu_{\text{opt}})$  on  $E_{\text{B}}$ , Extended Data Fig. 8c, d shows slices of  $\Delta I(\nu, \nu_{\text{opt}})$  at fixed values of  $h(\nu - \nu_{\text{2SB}}) = \pm 16$  meV, respectively, for three different binding energies  $E_{\text{B}} = 60$  meV (black curve), 240 meV (red curve) and 600 meV (blue curve). The slice at  $-16$  meV shows only one positive peak for  $\Delta_{\text{opt}} > 0$  and remains negative at  $\Delta_{\text{opt}} < 0$  for  $E_{\text{B}} = 60$  meV (black curve). Increasing the exciton binding energy gradually creates a new maximum within the region of negative detuning. This transition from a single- to a double-peaked structure can be exploited to assign the exciton binding energy as shown below. Monitoring the slices at  $h(\nu - \nu_{\text{2SB}}) = +16$  meV reveals a smooth, narrow, single-peaked structure close to  $\Delta_{\text{opt}} = 0$  for  $E_{\text{B}} = 60$  meV (black curve). By contrast, higher binding energies produce a more structured shape.

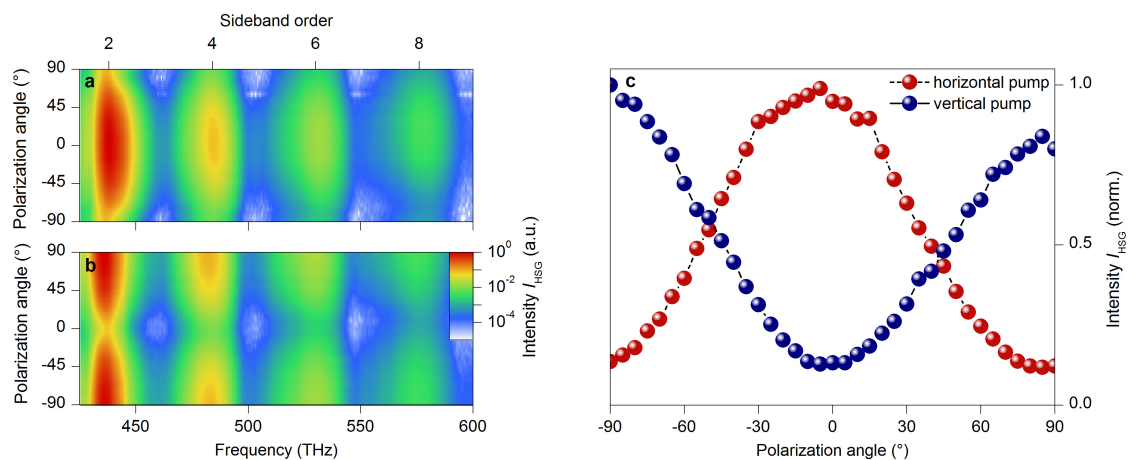
To construct differential spectra  $\Delta I(\nu, \nu_{\text{opt}})$  experimentally, we measure HSG spectra for narrowband pump pulses at different frequencies  $\nu_{\text{opt}}$  (see Fig. 1b, d) while varying  $t_{\text{ex}}$ . Extended Data Fig. 9c shows the resulting intensity map  $I_{\text{HSG}}(\nu, \nu_{\text{opt}}, t_{\text{ex}})$  for resonant excitation  $h\nu_{\text{opt}} = h\nu_{1s}$ , the corresponding computation is shown in Extended Data Fig. 9e. Both experiment and theory show the same spectral structure, including  $t_{\text{ex}}$ -dependent changes in the intensity and shape. A similar level of agreement is realized also for non-resonant excitations (not shown), paving the way to thorough theory–experiment comparison of the differential spectra.

On the basis of Extended Data Fig. 8, we expect the strongest dependence of  $\Delta I(\nu, \nu_{\text{opt}})$  on  $E_B$  for a finite detuning  $\Delta_{\text{opt}}$  from the 1s-exciton resonance. We therefore compare experimental traces  $\Delta I(\nu, \nu_{\text{opt}})$  at  $t_{\text{ex}} = -20$  fs for three different detunings  $\Delta_{\text{opt}} = -75$  meV (shaded area),  $-29$  meV (black curve) and  $27$  meV (red curve) in Extended Data Fig. 9a. A negative  $\Delta_{\text{opt}}$  produces a dispersive line shape to  $\Delta I(\nu, \nu_{\text{opt}})$ , that is, a dip followed by a zero crossing and a peak, which implies a shift of the underlying HSG intensity with a change in  $\nu_{\text{opt}}$ . For positive  $\Delta_{\text{opt}}$ , a double-peaked structure indicates a broadening of the HSG peaks. For a binding energy of  $E_B = 60$  meV, these line shapes are well reproduced by the theoretical computations using the same detuning  $\Delta_{\text{opt}}$  and excitation conditions as in the experiment, as shown in Extended Data Fig. 9b.

Extended Data Fig. 9d, f shows the corresponding calculated differentials  $\Delta I(\nu, \nu_{\text{opt}})$  when the exciton binding energy is increased to  $E_B = 240$  meV and  $E_B = 600$  meV, respectively. Whereas  $\Delta_{\text{opt}} = -29$  meV (black curve) and  $\Delta_{\text{opt}} = 27$  meV (red curve) produce a similar shape as for  $E_B = 60$  meV, the peak amplitude of the differential curve for  $\Delta_{\text{opt}} = -75$  meV (shaded area) greatly decreases for an intermediate binding energy of  $E_B = 240$  meV and turns into a negative dip for the largest binding energy. This structural change in  $\Delta I(\nu, \nu_{\text{opt}})$  can be used to deduce the exciton binding energy. We find that the experimental data are consistent with  $E_B = 60 \pm 20$  meV, which agrees well with values reported in literature<sup>30,46,47</sup>.

31. Sell, A., Leitenstorfer, A. & Huber, R. Phase-locked generation and field-resolved detection of widely tunable terahertz pulses with amplitudes exceeding 100 MV/cm. *Opt. Lett.* **33**, 2767–2769 (2008).
32. Gallot, G. & Grischkowsky, D. Electro-optic detection of terahertz radiation. *J. Opt. Soc. Am. B* **16**, 1204–1212 (1999).
33. Kira, M. & Koch, S. W. Many-body correlations and excitonic effects in semiconductor spectroscopy. *Prog. Quantum Electron.* **30**, 155–296 (2006).
34. Smith, R. P. *et al.* Extraction of many-body configurations from nonlinear absorption in semiconductor quantum wells. *Phys. Rev. Lett.* **104**, 247401 (2010).
35. Golde, D., Kira, M., Meier, T. & Koch, S. W. Microscopic theory of the extremely nonlinear terahertz response of semiconductors. *Phys. Status Solidi B* **248**, 863–866 (2011).
36. Danielson, J. R. *et al.* Interaction of strong single-cycle terahertz pulses with semiconductor quantum wells. *Phys. Rev. Lett.* **99**, 237401 (2007).
37. Girndt, A., Jahnke, F., Knorr, A., Koch, S. W. & Chow, W. W. Multi-band Bloch equations and gain spectra of highly excited II–VI semiconductor quantum wells. *Phys. Status Solidi B* **202**, 725–739 (1997).
38. Berger, C. *et al.* Novel type-II material systems for laser applications in the near-infrared regime. *ALP Advances* **5**, 047105 (2015).
39. Ramasubramanian, A. Large excitonic effects in monolayers of molybdenum and tungsten dichalcogenides. *Phys. Rev. B* **86**, 115409 (2012).
40. Yan, J.-Y. Theory of excitonic high-order sideband generation in semiconductors under a strong terahertz field. *Phys. Rev. B* **78**, 075204 (2008).
41. Vu, Q. T. *et al.* Light-induced gaps in semiconductor band-to-band transitions. *Phys. Rev. Lett.* **92**, 217403 (2004).
42. Vampa, G. *et al.* Theoretical analysis of high-harmonic generation in solids. *Phys. Rev. Lett.* **113**, 073901 (2014).
43. Banks, H. *et al.* Terahertz electron-hole recollisions in GaAs/AlGaAs quantum wells: robustness to scattering by optical phonons and thermal fluctuations. *Phys. Rev. Lett.* **111**, 267402 (2013).
44. Wang, H. *et al.* Transient nonlinear optical response from excitation induced dephasing in GaAs. *Phys. Rev. Lett.* **71**, 1261–1264 (1993).
45. Peyghambarian, N. *et al.* Blue shift of the exciton resonance due to exciton-exciton interactions in a multiple-quantum-well structure. *Phys. Rev. Lett.* **53**, 2433–2436 (1984).
46. Beal, A. R. & Liang, W. L. Excitons in 2H-WSe<sub>2</sub> and 3R-WSe<sub>2</sub>. *J. Phys. C* **9**, 2459–2466 (1976).
47. Mitioglu, A. A. *et al.* Optical investigation of monolayer and bulk tungsten diselenide (WSe<sub>2</sub>) in high magnetic fields. *Nano Lett.* **15**, 4387–4392 (2015).

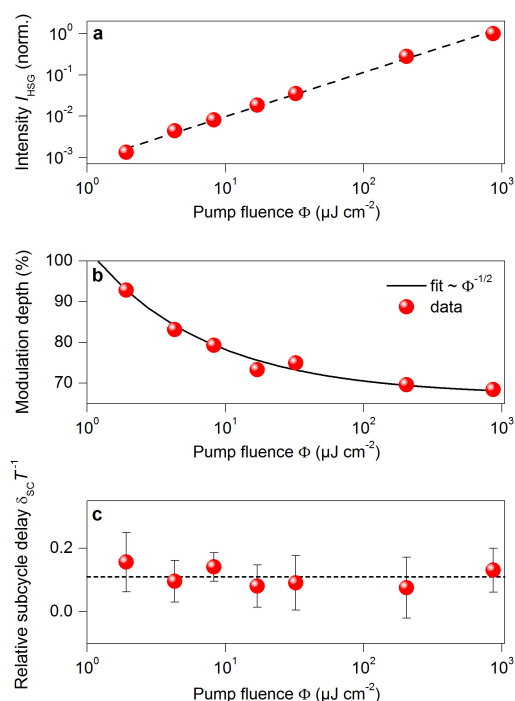




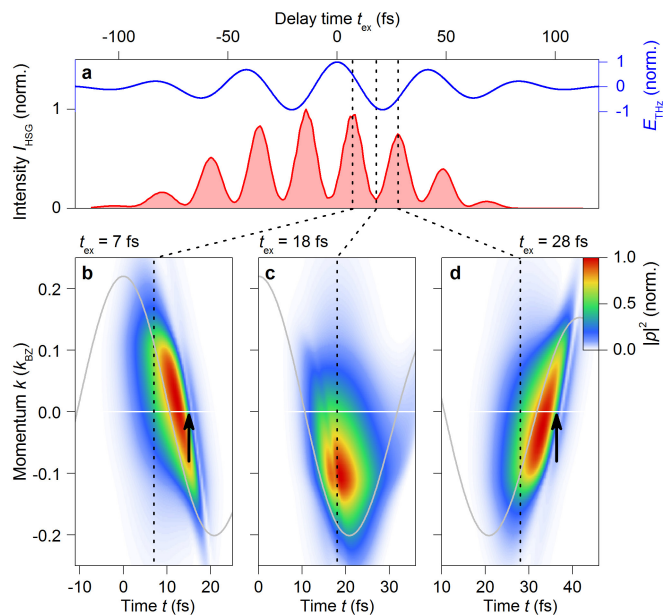
### Extended Data Figure 1 | Polarization of high-order sidebands.

**a, b,** False-colour plot of the spectral intensity of high-order sidebands (generated under resonant, spectrally narrow, optical excitation) of orders two to eight as a function of their polarization for horizontal (**a**) and

vertical (**b**) polarization of the interband excitation pulse. The polarization angle is defined such that  $0^\circ$  corresponds to a horizontally polarized excitation. **c,** Spectrally integrated sideband intensity as a function of the polarization angle.

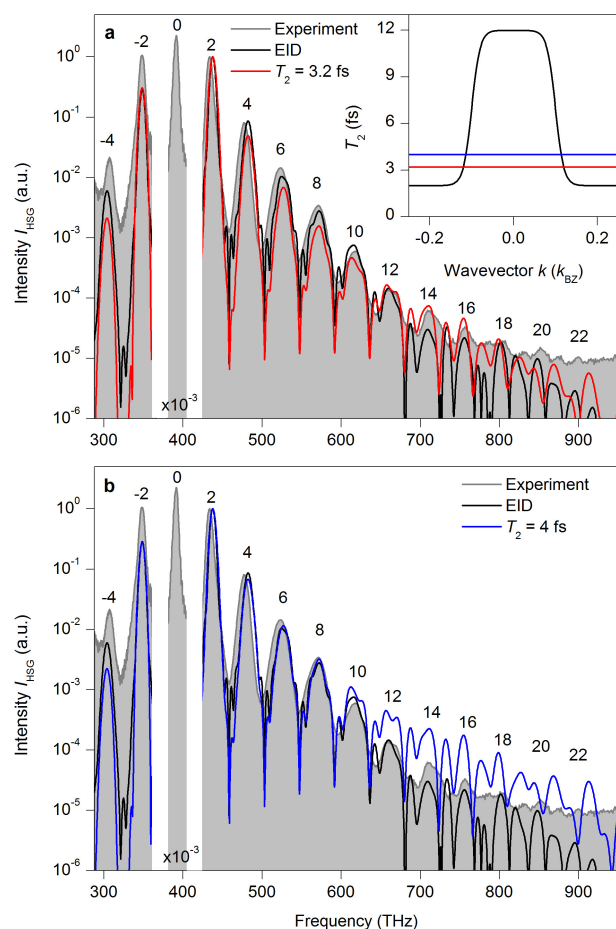


**Extended Data Figure 2 | Pump-fluence dependence of time-resolved high-order sideband generation.** **a**, The measured high-order sideband intensity  $I_{\text{HSG}}$  (spectrally and temporally integrated, red spheres) scales linearly with the pump fluence  $\Phi$ , as indicated by a guide to the eye (black dashed line). **b**, The modulation depth of the spectrally integrated temporal trace of  $I_{\text{HSG}}$  (red spheres, see Fig. 2b) decreases with increasing pump fluence, closely following a fit (black curve) proportional to the inverse square-root of the pump fluence. **c**, By contrast, the relative subcycle delay  $\delta_{\text{sc}} T^{-1}$  does not substantially change with increasing pump fluence. Red spheres represent the average of  $\delta_{\text{sc}} T^{-1}$  over the nine dominant, consecutive, high-order sideband peaks (see Fig. 2d) for different pump fluences; error bars indicate their standard deviation. The black dashed line marks the mean value of the displayed data points.

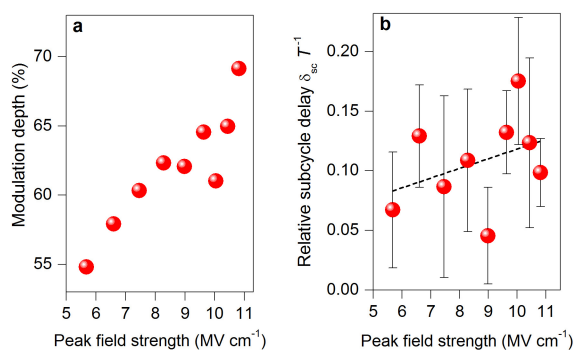


**Extended Data Figure 3 | Coherent excitonic polarization dynamics in  $k$  space.** **a**, Computed high-order sideband intensity  $I_{\text{HSG}}$  (red shaded area) and driving waveform (blue). Vertical black dotted lines highlight characteristic delay times  $t_{\text{ex}}$  at extrema of  $I_{\text{HSG}}$ . **b–d**, Coherent interband polarization  $|p|^2$  (colour scale) as a function of time  $t$  and reciprocal space coordinate  $k$  (in units of the wave vector  $k_{\text{BZ}}$  limiting the first Brillouin zone) for distinct delay times  $t_{\text{ex}}$  according to maximum (**b**,  $t_{\text{ex}} = 7$  fs; **d**,  $t_{\text{ex}} = 28$  fs) and minimum (**c**,  $t_{\text{ex}} = 18$  fs) HSG emission. Horizontal white lines mark  $k = 0$ . The driving field is depicted as grey curves and the recollision times (see Fig. 3d) are highlighted by black arrows.

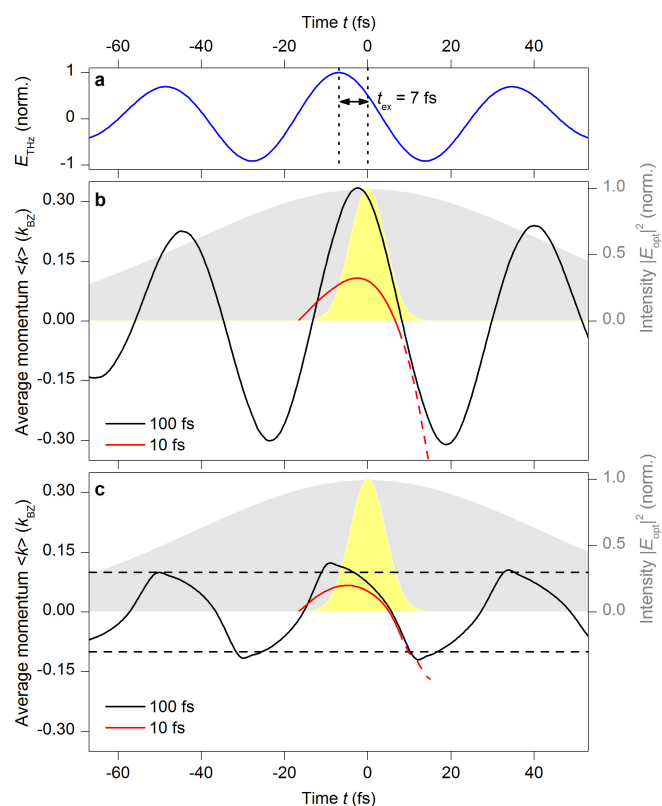




**Extended Data Figure 4 | Influence of dephasing on high-order sideband generation.** **a, b,** The measured sideband spectrum (shaded area) is compared with computations using constant dephasing times of  $T_2 = 3.2$  fs (**a**, red curve),  $T_2 = 4$  fs (**b**, blue curve) and a momentum-dependent dephasing model as presented in Fig. 1 (black curves; EID, excitation-induced dephasing). The sideband orders  $n$  are indicated above the relevant peaks. All spectra are normalized to the sideband peak corresponding to  $n = 2$ . The inset in **a** depicts the corresponding dephasing times  $T_2(k)$  as a function of the wave vector  $k$ ; the red (blue) horizontal line indicates a constant decay level  $T_2 = 3.2$  fs ( $T_2 = 4$  fs).



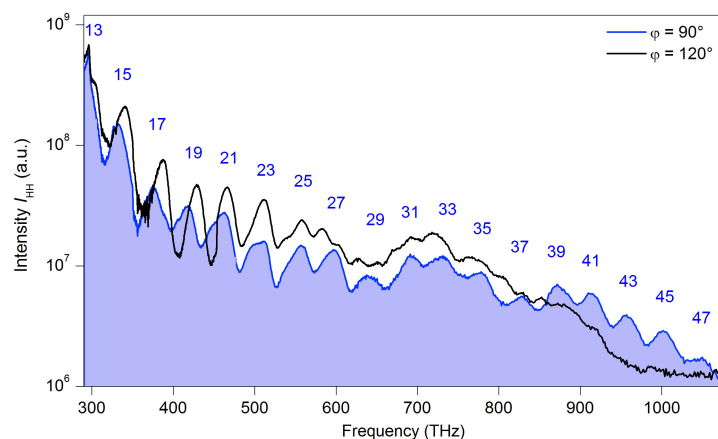
**Extended Data Figure 5 | Field scaling of time-resolved high-order sideband generation. a,** Scaling of the modulation depth (see discussion in Methods section ‘Influence of excitation fluence’) of temporal traces of high-order sideband intensity with the driving peak field strength. An increase in the modulation depth suggests a faster dephasing  $T_2$ . **b,** Subcycle delay  $\delta_{sc}$  in units of the driving period  $T$  averaged over the eight most-dominant sideband peaks as a function of the peak field. The error bars represent the standard deviation of the eight peaks and the dashed line depicts a linear fit to the data points.



#### Extended Data Figure 6 | Influence of excitation-induced dephasing.

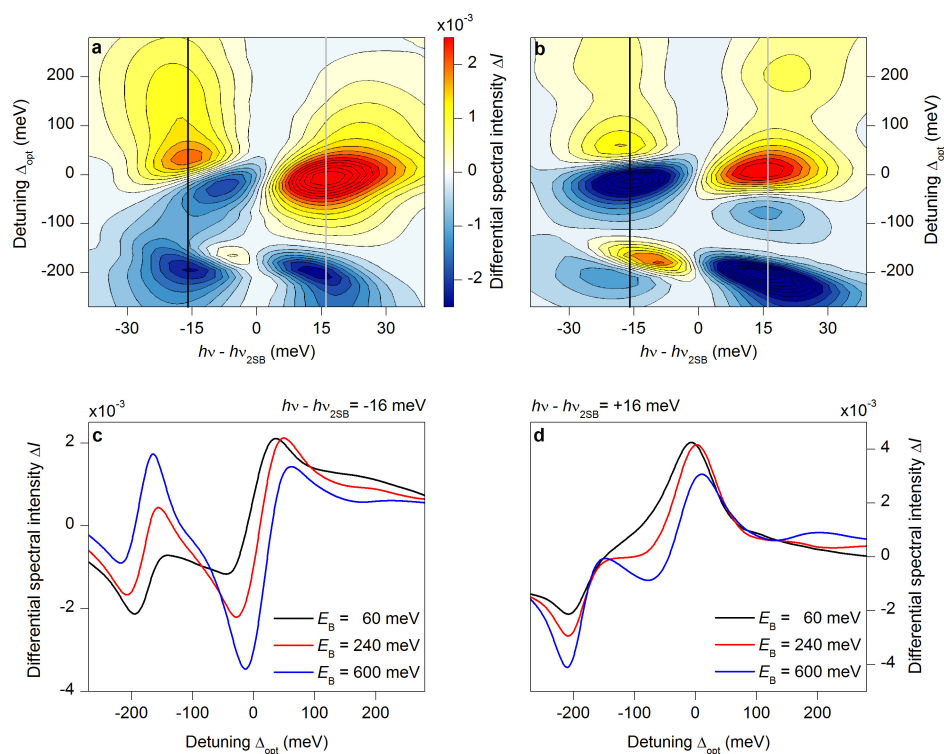
**a**, Terahertz waveform used in the computations. **b**, **c**, Mean electron excursion  $\langle k \rangle$  calculated with a constant dephasing  $T_2 = 12$  fs (**b**) and excitation-induced dephasing (**c**) for a delay of  $t_{\text{ex}} = 7$  fs (see vertical dotted lines in **a**). The shaded areas indicate the intensity envelopes of the excitation pulse with a full-width at half-maximum of 10 fs (yellow) and 100 fs (grey). The corresponding mean momenta  $\langle k \rangle$  are shown as red and black curves, respectively. The red dashed curves show  $\langle k \rangle$  after the electron-hole collision. The dashed horizontal lines in **c** mark the positions in  $k$  space where the scattering time  $T_2(k)$  switches from slow to fast dephasing.





**Extended Data Figure 7 | High-order harmonic generation in tungsten diselenide.** An intensity spectrum (blue shaded area) of THz-driven high-order harmonic generation in 60-nm-thick WSe<sub>2</sub> shows distinct peaks at odd orders  $n' = 13$  to  $n' = 47$  (indicated by numerals) of the fundamental

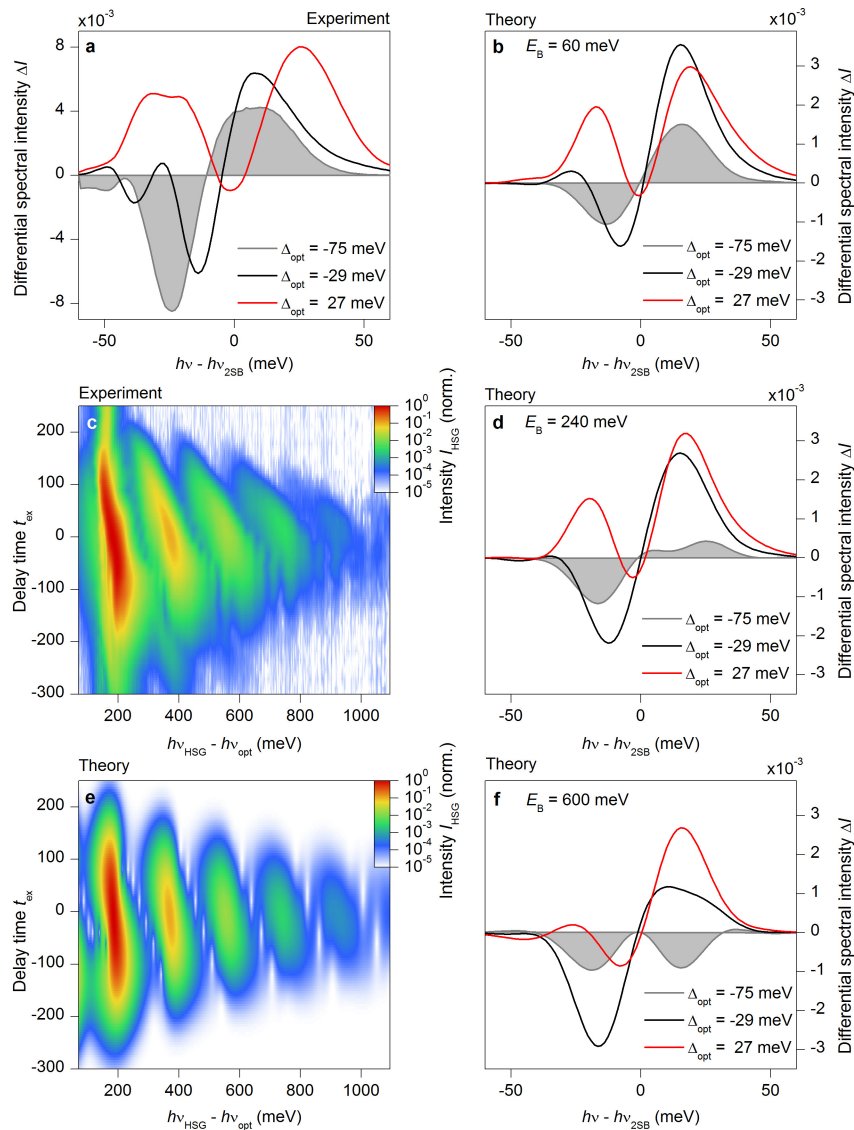
frequency  $\nu_{\text{THz}} = 22.3$  THz (peak electric field in air,  $21 \text{ MV cm}^{-1}$ ;  $\varphi = 90^\circ$ ). An intensity spectrum for  $\varphi = 120^\circ$  is also shown as a black curve. The spectral intensity has been corrected for the grating efficiency and for the quantum efficiency of the spectrograph used.



### Extended Data Figure 8 | Differential sideband spectroscopy.

**a, b,** Contours of differential spectra  $\Delta I(\nu, \nu_{\text{opt}})$  (colour scale) as a function of  $h(\nu - \nu_{2\text{SB}})$  and detuning  $\Delta_{\text{opt}}$  computed for an excitonic binding energy of  $E_B = 60$  meV (**a**) and  $E_B = 600$  meV (**b**). The black and grey vertical lines mark the positions of the slices shown in **c** and **d**,

respectively. **c, d,** Snapshots of  $\Delta I(\nu, \nu_{\text{opt}})$  at fixed values of  $|h(\nu - \nu_{2\text{SB}})| = 16$  meV below (**c**) and above (**d**) the second-order sideband peak for three different binding energies  $E_B = 60$  meV (black curves),  $E_B = 240$  meV (red curves) and  $E_B = 600$  meV (blue curves).



**Extended Data Figure 9 | Quantitative analysis of the binding energy.** **a**, Measured differential spectra  $\Delta I(\nu, \nu_{\text{opt}})$  for three different detunings  $\Delta_{\text{opt}} = -75$  meV (shaded area),  $\Delta_{\text{opt}} = -29$  meV (black curve) and  $\Delta_{\text{opt}} = 27$  meV (red curve) as functions of  $\nu$ , centred at the position of the second sideband  $\nu_{2\text{SB}}$ . **b, d, f**, Computed differential spectra  $\Delta I(\nu, \nu_{\text{opt}})$  for

binding energies of  $E_B = 60$  meV (**b**),  $E_B = 240$  meV (**d**) and  $E_B = 600$  meV (**f**) and detunings  $\Delta_{\text{opt}}$  as in the experiment shown in **a**. **c, e**, Measured (**c**) and calculated (**e**) HSG intensities (colour scale) for  $\Delta_{\text{opt}} = 0$  as functions of  $\nu_{\text{HSG}}$  and delay  $t_{\text{ex}}$ .



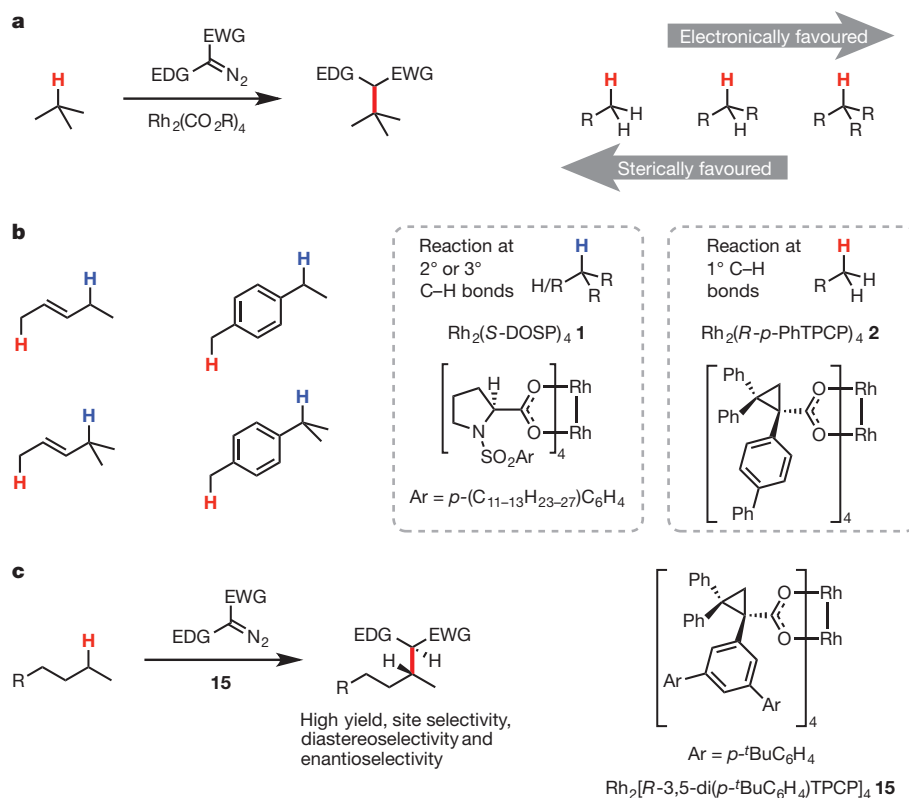
# Site-selective and stereoselective functionalization of unactivated C–H bonds

Kuangbiao Liao<sup>1</sup>, Solymar Negretti<sup>1</sup>, Djamaladdin G. Musaev<sup>2</sup>, John Bacsá<sup>1</sup> & Huw M. L. Davies<sup>1</sup>

The laboratory synthesis of complex organic molecules relies heavily on the introduction and manipulation of functional groups, such as carbon–oxygen or carbon–halogen bonds; carbon–hydrogen bonds are far less reactive and harder to functionalize selectively. The idea of C–H functionalization, in which C–H bonds are modified at will instead of the functional groups, represents a paradigm shift in the standard logic of organic synthesis<sup>1–3</sup>. For this approach to be generally useful, effective strategies for site-selective C–H functionalization need to be developed. The most practical solutions to the site-selectivity problem rely on either intramolecular reactions<sup>4</sup> or the use of directing groups within the substrate<sup>5–8</sup>. A challenging, but potentially more flexible approach, would be to use catalyst control to determine which site in a particular substrate would be functionalized<sup>9–11</sup>. Here we describe the use of dirhodium catalysts to achieve highly site-selective,

diastereoselective and enantioselective C–H functionalization of *n*-alkanes and terminally substituted *n*-alkyl compounds. The reactions proceed in high yield, and functional groups such as halides, silanes and esters are compatible with this chemistry. These studies demonstrate that high site selectivity is possible in C–H functionalization reactions without the need for a directing or anchoring group present in the molecule.

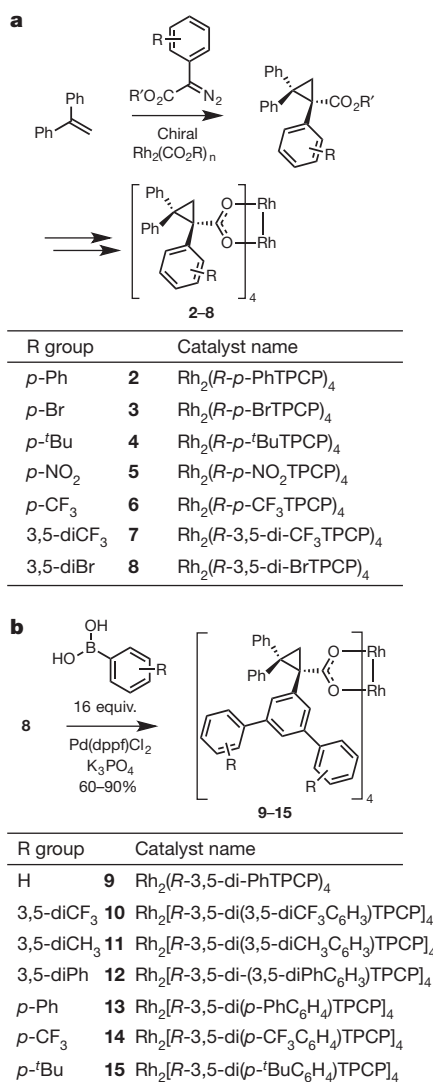
We have demonstrated that catalyst-controlled site selectivity is possible in the dirhodium-catalysed intermolecular reactions of donor/acceptor carbenes with relatively activated C–H bonds, such as those at benzylic and allylic positions<sup>12</sup> and  $\alpha$  to oxygen<sup>13</sup>. The carbene-induced C–H functionalization is initiated by a hydride transfer event, and consequently the reaction is favoured at sites capable of stabilizing a build-up of positive charge; thus, tertiary C–H bonds are electronically preferred (Fig. 1a)<sup>14</sup>. However, the dirhodium–carbene



**Figure 1 | Site-selective C–H functionalization by donor/acceptor carbenes.** **a**, Site-selectivity trend in dirhodium-catalysed C–H functionalization by donor/acceptor carbenes is controlled by a delicate balance between steric and electronic effects. **b**, Previous studies showed that site selectivity at activated C–H bonds could be controlled by the nature of the catalyst.  $\text{Rh}_2(\text{S-DOSP})_4$  prefers functionalization at

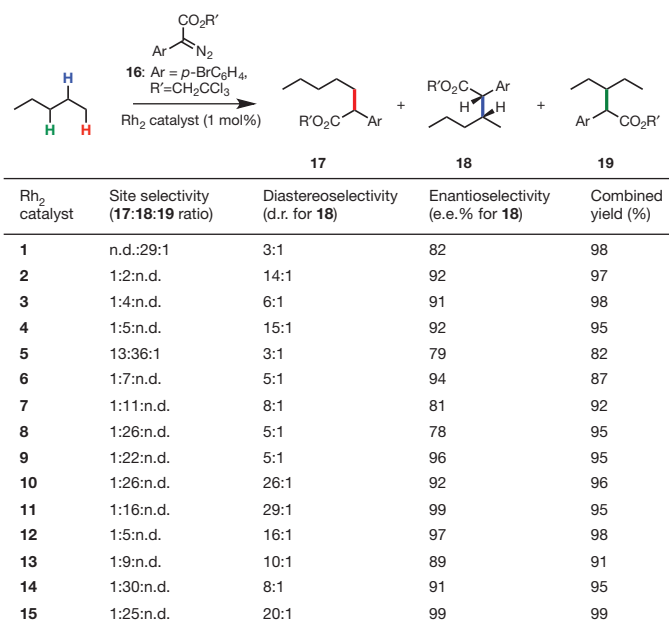
secondary and tertiary C–H bonds, whereas  $\text{Rh}_2(\text{R-p-PhTPCP})_4$  prefers functionalization at primary C–H bonds. **c**, This study demonstrates that highly site-selective and stereoselective C–H functionalization of unactivated C–H bonds can be achieved with the new catalyst  $\text{Rh}_2[\text{R-3,5-di}(p\text{-}^t\text{BuC}_6\text{H}_4)\text{TPCP}]_4$ .

<sup>1</sup>Department of Chemistry, Emory University, 1515 Dickey Drive, Atlanta, Georgia 30322, USA. <sup>2</sup>Cherry L. Emerson Center for Scientific Computation, Emory University, 1521 Dickey Drive, Atlanta, Georgia, 30322, USA.



**Figure 2 | Synthesis of TPCP carboxylate dirhodium catalysts.** **a**, Original synthesis of TPCP carboxylate dirhodium catalysts. **b**, New synthesis of TPCP carboxylate dirhodium catalysts using palladium-catalysed cross-coupling on a preformed dirhodium complex. See Supplementary Information for the synthetic details.

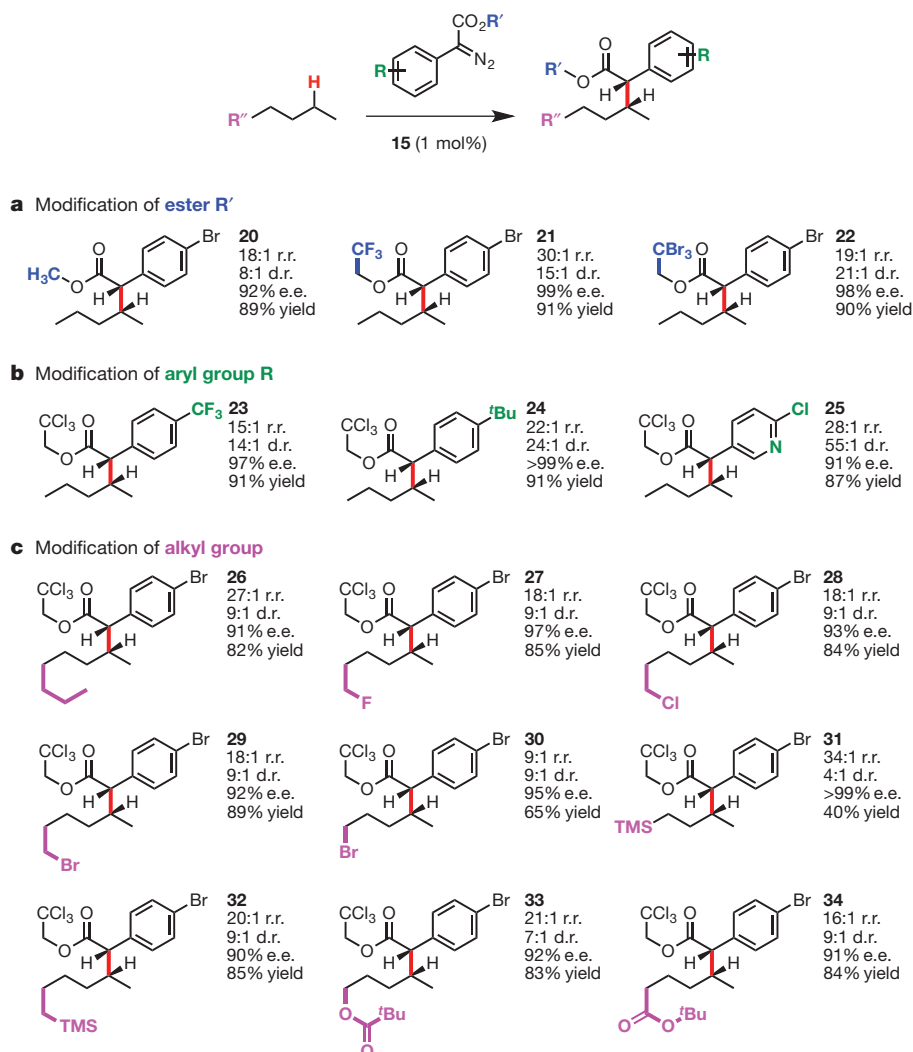
complex is sterically demanding, so on steric grounds the primary C–H bond would be preferred. We have shown that by altering the steric nature of the catalyst, the site selectivity of C–H functionalization can be dictated in these activated substrates from a preference for the secondary or tertiary C–H bonds using the established dirhodium catalyst Rh<sub>2</sub>(S-DOSP)<sub>4</sub> (**1**), to the primary C–H bond with the more sterically encumbered triaryl cyclopropanecarboxylate catalyst Rh<sub>2</sub>(*R*-*p*-PhTPCP)<sub>4</sub> (**2**) (Fig. 1b)<sup>12,13</sup>. Even though intermolecular C–H functionalization with metal carbene intermediates has been a longstanding challenge<sup>4,9,14–16</sup>, the metal-catalysed reactions of *n*-alkanes with the more traditional acceptor carbene derived from ethyl diazoacetate has been explored, furnishing mixtures of products, controlled to some extent by the nature of the catalyst, with primary C–H insertion preferred when employing larger and more electrophilic catalysts<sup>9,15</sup>. The reaction of *n*-alkanes with the donor/acceptor carbene derived from methyl phenyldiazoacetate showed even greater preference for primary C–H insertion when a bulky rhodium porphyrin catalyst was used<sup>17</sup>. We describe a new triaryl cyclopropanecarboxylate catalyst, Rh<sub>2</sub>[*R*-3,5-di(*p*-<sup>t</sup>BuC<sub>6</sub>H<sub>4</sub>)TPCP]<sub>4</sub> (**15**), that can achieve highly regio-, diastereo- and enantioselective C–H functionalization at the unactivated C2 position of *n*-alkanes or terminally substituted *n*-alkyl compounds (Fig. 1c).



**Figure 3 | Catalyst optimization studies.** Evaluation of the dirhodium TPCP catalysts for the C–H functionalization of pentane. The optimum catalyst for high site selectivity at C2, diastereoselectivity, enantioselectivity and yield is catalyst **15**, Rh<sub>2</sub>[*R*-3,5-di(*p*-<sup>t</sup>BuC<sub>6</sub>H<sub>4</sub>)TPCP]<sub>4</sub>. The standard reaction was carried out in refluxing pentane with 1 mol% catalyst loading. See Supplementary Information for experimental details. n.d., not determined.

Having discovered that the triphenylcyclopropane carboxylate (TPCP) catalysts can exert considerable influence on the site selectivity of donor/acceptor carbenes with activated substrates for C–H functionalization<sup>12,13</sup>, we decided to explore whether these catalysts could achieve selective intermolecular C–H functionalization of unactivated C–H bonds. To initiate this project, we prepared a series of chiral dirhodium catalyst derivatives, **2–8**, was prepared by our standard method, which involves asymmetric cyclopropanation, hydrolysis of the resulting ester to the carboxylic acid, enantiomer enrichment, and ligand exchange (Fig. 2a)<sup>12</sup>. When it became clear that catalysts **7** and **8** with a 3,5-disubstituted phenyl group had promising properties (see later), we developed a new approach for catalyst diversification by conducting an eightfold palladium-catalysed cross-coupling reaction on the fully formed 3,5-dibromo complex **8** to generate a second series of catalysts (**9–15**) (Fig. 2b). The new approach has broad utility because the catalysts are made from a single enantiomerically pure complex, avoiding the variable yields observed in ligand exchange reactions when the ligands become sterically congested and streamlining the whole process.

The initial evaluation of the different catalysts was conducted using *n*-pentane as the test substrate. Even though *n*-pentane is a simple substrate, the challenge of C–H functionalization of this system is apparent when one considers the subtlety required in achieving a site-selective reaction. Recent studies have shown that donor/acceptor carbene reactions tend to proceed better when the trichloroethyl ester is used instead of the methyl ester<sup>13</sup>. Hence, 2,2,2-trichloroethyl 2-(4-bromophenyl)-2-diazoacetate (**16**) was used as the carbene precursor for the initial evaluation (Fig. 3). The *N*-sulfonylproline Rh<sub>2</sub>(*R*-DOSP)<sub>4</sub> (**1**) is the standard chiral catalyst that has been used for most of the C–H functionalization chemistry of donor/acceptor carbenes, but its reaction with *n*-alkanes has not been reported<sup>14,16</sup>. Therefore, it was promising to observe that when Rh<sub>2</sub>(*R*-DOSP)<sub>4</sub> was used as the catalyst, **16** reacted with pentane to give only **18** and **19**. The products derived from insertion into methylene C–H bonds were observed in a high overall yield and good site selectivity (29:1



**Figure 4** | C–H functionalization of alkanes and substituted *n*-alkanes. **a–c**, Evaluation of the scope of  $\text{Rh}_2[\text{R}-3,5\text{-di}(p\text{-}^i\text{BuC}_6\text{H}_4)\text{TPCP}]_4$ -catalysed C–H functionalization of pentane and substituted *n*-alkanes. **a**, Compounds 20–22 illustrate the effect of the ester group. **b**, Compounds 23–25 illustrate the effect of modifications to the donor

group. **c**, Compounds 26–34 illustrate the effect of the alkyl substituent. Reactions with pentane were conducted in pentane as solvent. Reactions with substituted alkanes were conducted with 3 equiv. of the substituted alkane with dichloromethane as solvent. See Supplementary Information for complete experimental details.

regioisomeric ratio (r.r.) of C2 to C3). The enantiocontrol was also reasonable (82% enantiomeric excess (e.e.)), but the diastereocontrol was moderate (3:1 diastereomeric ratio (d.r.)). Even though the reaction displayed exceptional site selectivity considering that both reacting sites are methylene sites with only a slight difference in steric environment, we recognized that the competition with the internal methylene sites would probably lead to complex mixtures when longer *n*-alkanes were used as substrates. Therefore, we examined the reaction with the triarylcyclopropanecarboxylate catalysts, as previous studies had shown that these catalysts behave as more sterically encumbered catalysts and drive the C–H functionalization site selectivity towards the less sterically hindered positions<sup>12,13</sup>. This expected trend was followed by the *p*-substituted catalysts 2–6, which generated a considerable amount of the C1 insertion product 17 in addition to C2 18 (1:3 to 1:7 ratio), without any C3 insertion product 19 being formed. The diastereo- and/or enantioselectivity of the C–H functionalization with catalysts 2–6 is influenced by the nature of the aryl substituent on the ligand, with the *p*-Ph derivative 2 giving the highest level of enantioselectivity (92% e.e.) and the *p*-*tert*-butyl derivative 4 giving the highest diastereoselectivity (15:1 d.r.). When the reaction was extended to the 3,5-di- $\text{CF}_3$ -substituted catalyst 7, we expected increased formation of the C1 insertion product. However, this was not the case, and a strong preference for C2 over C1 functionalization (11:1 ratio) was

observed without any formation of the C3 functionalized product 19. Furthermore, the diastereoselectivity and enantioselectivity for the formation of 18 was quite high (8:1 d.r., 81% e.e.). A similar effect was also seen with the 3,5-di-Br-substituted catalyst 8, which gave an even higher C2 to C1 ratio (26:1), but slightly lower levels of diastereoselectivity and enantioselectivity (5:1 d.r., 78% e.e.). On the basis of these results, the study was extended to the 3,5-diaryl-substituted catalysts 9–15, prepared from the cross-coupling strategy on the dirhodium complex 8. Many of these catalysts gave a strong preference for C–H functionalization at C2 over C1 (up to 30:1), some generated the C2 product 18 with exceptional diastereoselectivity (up to 29:1 d.r.), and all gave high levels of enantioselectivity (89–99% e.e.). The optimum catalyst in terms of overall performance was the 3,5-di(*p*-*tert*-butylphenyl)phenyl-derived catalyst 15,  $\text{Rh}_2[\text{R}-3,5\text{-di}(p\text{-}^i\text{BuC}_6\text{H}_4)\text{TPCP}]_4$ , which gave the C2 product 18 in 99% overall yield with high site selectivity favouring C2 over C1 (25:1 ratio), diastereoselectivity (20:1 d.r.) and enantioselectivity (99% e.e.).

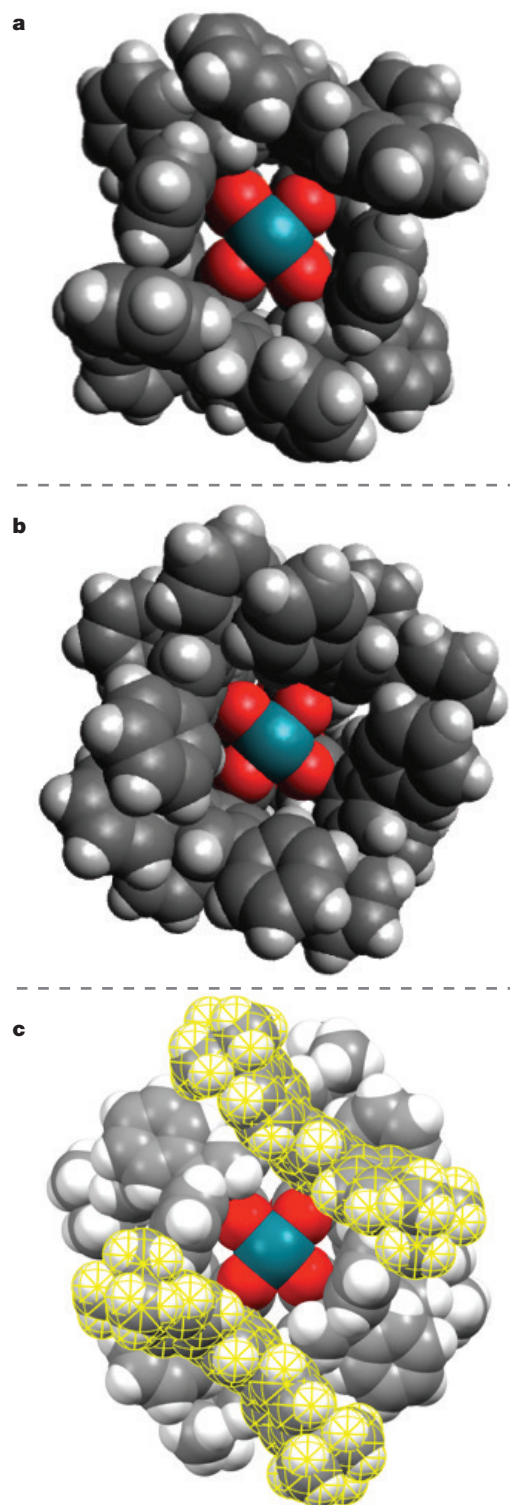
The  $\text{Rh}_2[\text{R}-3,5\text{-di}(p\text{-}^i\text{BuC}_6\text{H}_4)\text{TPCP}]_4$ -catalysed reactions with pentane were then examined with a variety of donor/acceptor carbene precursors (Fig. 4a, b). All of the reactions proceeded in high yield (87–91%) and with high levels of enantioselectivity (91% to >99% e.e.). The first three substrates were chosen to examine the nature of the ester substituent (Fig. 4a). The methyl ester derivative



was quite effective in the C–H functionalization to form **20**, but overall its performance in terms of yield and selectivity was lower than the trichloroethyl ester **16**. Larger alkyl esters are best avoided when attempting intermolecular C–H functionalization reactions, as intramolecular C–H functionalization will tend to compete. However, the trifluoroethyl and tribromoethyl esters are effective ester groups and perform comparatively to the trichloroethyl derivatives to form **21** and **22**. In specific cases, the trifluoroethyl group may be worth considering further because the site selectivity for C2 for **21** is better than with the trichloroethyl derivative (30:1 versus 25:1 r.r.), although the diastereoselectivity is lower (15:1 versus 20:1 d.r.). The reaction was extended to a series of phenyl-substituted carbenes (Fig. 4b). They all performed well in these reactions, but the site selectivity and diastereoselectivity were somewhat reduced for the *p*-trifluoromethyl derivative **23** (14:1 C2 versus C1, 15:1 d.r.). The potential breadth of utility of the C–H functionalization is illustrated with a pyridine system, which is an effective substrate for this chemistry. Indeed, the C–H functionalization product **25** was produced with the highest diastereoselectivity to date (55:1 d.r.).

As the selectivity competition is between the C2 and C1 positions, the ratio would be expected to remain about the same for longer-chain alkanes. This was indeed the case for the reaction with *n*-octane, which generated **26** with a site selectivity of 27:1 r.r., slightly enhanced compared with pentane (Fig. 4c). One of the challenges of C–H functionalization is to conduct the reactions in the presence of functional groups. Having established that  $\text{Rh}_2[\text{R-3,5-di}(p\text{-}^t\text{BuC}_6\text{H}_4)\text{TPCP}]_4$  is an exceptional catalyst for C–H functionalization of *n*-pentane, we performed an initial evaluation to determine if the reaction can be conducted in the presence of terminally substituted *n*-alkyl compounds, such as alkyl halides, silanes and esters (Fig. 4c). These reactions were conducted with the bromophenyl derivative **16** and 3 equiv. of the substrate in dichloromethane as solvent under reflux conditions. Alkyl halides are versatile functionality in organic synthesis but as the dirhodium carbene is highly electrophilic we considered whether they would be compatible with the C–H functionalization chemistry. Indeed, 1-bromo-, 1-chloro- and 1-fluorohexane were found to be good substrates for the C–H functionalization. In all three cases, the levels of enantioselectivity for the formation of **27–29** were high (92–97% e.e.), the site selectivity for the methylene C–H bond was 18:1 but the diastereoselectivity was somewhat diminished (9:1 d.r.). The reaction with 1-bromopentane to form **30** also proceeded with high enantioselectivity (95% e.e.) but the yield was only 65% and the site selectivity was only 9:1 r.r. This result may indicate that the bromine is displaying a long-range inductive effect, which slightly deactivates the C2 position. Such a characteristic could become a useful controlling element in more complex systems. *n*-Alkyl silanes were considered to be interesting substrates because electronically the C–H bonds  $\beta$  to silicon should be electronically activated towards C–H functionalization. The steric influence dominates in these reactions, and preferred formation of the regioisomers **31** and **32** was observed once again. Another functional group that is compatible with this chemistry is an ester, as illustrated in the formation of **33** and **34**, again with strong preference for functionalization of the methylene C–H bond. These studies demonstrate that the  $\text{Rh}_2[\text{R-3,5-di}(p\text{-}^t\text{BuC}_6\text{H}_4)\text{TPCP}]_4$ -catalysed reactions of donor/acceptor carbenes have a strong preference for functionalization of the methylene site at the C2 position of alkanes and terminally substituted alkanes. Presumably, the C–H functionalization at the methyl group is less favoured on electronic grounds, whereas the steric environment around the catalyst is sufficient to distinguish between the methylene sites. All of the reactions proceed with high enantioselectivity (90% to >99% e.e.), but the diastereoselectivity is variable (4:1–9:1 d.r.).

Further evidence to help understand why  $\text{Rh}_2[\text{R-3,5-di}(p\text{-}^t\text{BuC}_6\text{H}_4)\text{TPCP}]_4$  is such an effective catalyst was obtained from computational and X-ray crystallographic studies. We and others have previously shown that the presence of four identical chiral ligands around the



**Figure 5 | Structural information about the dirhodium catalysts.** **a**, Computational structure of the  $\alpha, \beta, \alpha, \beta$  form of  $\text{Rh}_2[\text{S-3,5-diPhTPCP}]_4$ . **b**, Computational structure of the  $\alpha, \alpha, \alpha, \alpha$  form of  $\text{Rh}_2(\text{S-3,5-diPhTPCP})_4$  (5.0 kcal mol<sup>−1</sup> less stable than the  $\alpha, \beta, \alpha, \beta$  form). **c**, X-ray crystal structure of  $\text{Rh}_2[\text{S-3,5-di}(p\text{-}^t\text{BuC}_6\text{H}_4)\text{TPCP}]_4$ , in which the two 3,5-di(*p*-<sup>t</sup>BuC<sub>6</sub>H<sub>4</sub>)C<sub>6</sub>H<sub>3</sub>- groups on the top face are highlighted in yellow. For clarity, the diethyl ether molecules that were coordinated to the rhodium in the crystal structure have been removed. See the Supplementary Information for complete experimental details.

dirhodium core can result in a catalyst with higher symmetry than the ligands themselves<sup>18–22</sup>. Computational studies on  $\text{Rh}_2[\text{S-3,5-diPhTPCP}]_4$  revealed that the 3,5-disubstituted pattern disfavors



the existence of two adjacent 3,5-diphenylphenyl groups on the same face of the dirhodium catalyst. Hence the preferred orientation of the catalyst is the  $\alpha, \beta, \alpha, \beta$ -orientation (Fig. 5a), which is favoured over the  $\alpha, \alpha, \alpha, \alpha$ -orientation (Fig. 5b) by 5.0 kcal mol<sup>-1</sup>. The results from our computational studies were consistent with data from X-ray crystallographic analysis. An X-ray crystal structure of **15** as the corresponding bis(diethyl ether) complex was solved, and the complex was found to adopt the  $\alpha, \beta, \alpha, \beta$ -orientation (Fig. 5c). This arrangement is D<sub>2</sub> symmetric, which causes both faces of the dirhodium catalyst to be the same, and limits the number of possible orientations of the carbene when it binds to the dirhodium core. The C–H functionalization has been proposed to proceed via a concerted asynchronous mechanism in which the hydrogen of the C–H bond first approaches the carbene site<sup>23</sup>. It has been previously demonstrated that when C–H insertion into methylene C–H bonds occurs, the reactions can be highly diastereoselective as long as there is considerable size differentiation between the two other groups at the methylene site<sup>14</sup>. The size difference between a methyl and an *n*-propyl group, as in the case of pentane, would not be expected to be sufficient to cause high levels of diastereoselectivity using the more traditional chiral dirhodium tetracarboxylate catalysts, but when the bulky TPCP catalysts are used, high diastereoselectivity is also possible.

These studies demonstrate that highly site-selective C–H functionalization of unactivated C–H bonds at C2 of alkanes and terminally substituted alkanes is a viable process. The reactions proceed in high yield and functional groups such as halides, silanes and esters are compatible with this chemistry. A new class of D<sub>2</sub>-symmetric dirhodium catalysts were developed that are capable of achieving not only high site selectivity in these C–H functionalization reactions, but also high levels of diastereocontrol and enantiocontrol. These results show that high site selectivity is possible in C–H functionalization reactions without the need for a directing or anchoring group present in the molecule. The demonstration that it is possible to design catalysts with defined chiral pockets to control not only the enantioselectivity, but also the diastereo- and site selectivity of carbene-induced C–H functionalization, could have broad implications on future research directions on selective C–H functionalization.

Received 13 January; accepted 8 March 2016.

- Yamaguchi, J., Yamaguchi, A. D. & Itami, K. C–H bond functionalization: emerging synthetic tools for natural products and pharmaceuticals. *Angew. Chem. Int. Ed.* **51**, 8960–9009 (2012).
- Gutekunst, W. R. & Baran, P. S. C–H functionalization logic in total synthesis. *Chem. Soc. Rev.* **40**, 1976–1991 (2011).
- Hartwig, J. F. Evolution of C–H bond functionalization from methane to methodology. *J. Am. Chem. Soc.* **138**, 2–24 (2016).
- Doyle, M. P., Duffy, R., Ratnikov, M. & Zhou, L. Catalytic carbene insertion into C–H bonds. *Chem. Rev.* **110**, 704–724 (2010).
- Zhang, F. & Spring, D. R. Arene C–H functionalisation using a removable/modifiable or a traceless directing group strategy. *Chem. Soc. Rev.* **43**, 6906–6919 (2014).
- Topczewski, J. J. & Sanford, M. S. Carbon–hydrogen (C–H) bond activation at Pd<sup>IV</sup>: a frontier in C–H functionalization catalysis. *Chem. Sci.* **6**, 70–76 (2015).
- Engle, K. M., Mei, T.-S., Wasa, M. & Yu, J.-Q. Weak coordination as a powerful means for developing broadly useful C–H functionalization reactions. *Acc. Chem. Res.* **45**, 788–802 (2012).
- Colby, D. A., Bergman, R. G. & Ellman, J. A. Rhodium-catalyzed C–C bond formation via heteroatom-directed C–H bond activation. *Chem. Rev.* **110**, 624–655 (2010).
- Caballero, A. *et al.* Catalytic functionalization of low reactive C(sp<sup>3</sup>)-H and C(sp<sup>2</sup>)-H bonds of alkanes and arenes by carbene transfer from diazo compounds. *Dalton Trans.* **44**, 20295–20307 (2015).
- Kuhl, N., Hopkinson, M. N., Wencel-Delord, J. & Glorius, F. Beyond directing groups: transition-metal-catalyzed C–H activation of simple arenes. *Angew. Chem. Int. Ed.* **51**, 10236–10254 (2012).
- Mkhalid, I. A. I., Barnard, J. H., Marder, T. B., Murphy, J. M. & Hartwig, J. F. C–H activation for the construction of C–B bonds. *Chem. Rev.* **110**, 890–931 (2010).
- Qin, C. M. & Davies, H. M. L. Role of sterically demanding chiral dirhodium catalysts in site-selective C–H functionalization of activated primary C–H bonds. *J. Am. Chem. Soc.* **136**, 9792–9796 (2014).
- Guptill, D. M. & Davies, H. M. L. 2,2,2-Trichloroethyl aryldiazoacetates as robust reagents for the enantioselective C–H functionalization of methyl ethers. *J. Am. Chem. Soc.* **136**, 17718–17721 (2014).
- Davies, H. M. L. & Morton, D. Guiding principles for site selective and stereoselective intermolecular C–H functionalization by donor/acceptor rhodium carbenes. *Chem. Soc. Rev.* **40**, 1857–1869 (2011).
- Demonceau, A., Noels, A. F., Hubert, A. J. & Teyssie, P. Transition-metal-catalyzed reactions of diazoesters—insertion into C–H bonds of paraffins by carbenoids. *J. Chem. Soc. Chem. Commun.* **14**, 688–689 (1981).
- Davies, H. M. L., Hansen, T. & Churchill, M. R. Catalytic asymmetric C–H activation of alkanes and tetrahydrofuran. *J. Am. Chem. Soc.* **122**, 3063–3070 (2000).
- Thu, H.-Y. *et al.* Highly selective metal catalysts for intermolecular carbenoid insertion into primary C–H bonds and enantioselective C–C bond formation. *Angew. Chem. Int. Ed.* **47**, 9747–9751 (2008).
- Hansen, J. & Davies, H. M. L. High symmetry dirhodium(II) paddlewheel complexes as chiral catalysts. *Coord. Chem. Rev.* **252**, 545–555 (2008).
- Davies, H. M. L., Bruzinski, P. R., Lake, D. H., Kong, N. & Fall, M. J. Asymmetric cyclopropanations by rhodium(II) *N*-(arylsulfonyl)proline catalyzed decomposition of vinyl diazomethanes in the presence of alkenes. Practical enantioselective synthesis of the four stereoisomers of 2-phenylcyclopropan-1-amino acid. *J. Am. Chem. Soc.* **118**, 6897–6907 (1996).
- Qin, C. M. *et al.* D<sub>2</sub>-symmetric dirhodium catalyst derived from a 1,2,2-triarylcyclopropanecarboxylate ligand: design, synthesis and application. *J. Am. Chem. Soc.* **133**, 19198–19204 (2011).
- DeAngelis, A., Dmitrenko, O., Yap, G. P. A. & Fox, J. M. Chiral crown conformation of Rh<sub>2</sub>(S-PTTL)<sub>4</sub>: Enantioselective cyclopropanation with  $\alpha$ -alkyl- $\alpha$ -diazoesters. *J. Am. Chem. Soc.* **131**, 7230–7231 (2009).
- Lindsay, V. N. G., Lin, W. & Charette, A. B. Experimental evidence for the all-up reactive conformation of chiral rhodium(II) carboxylate catalysts: enantioselective synthesis of *cis*-cyclopropane  $\alpha$ -amino acids. *J. Am. Chem. Soc.* **131**, 16383–16385 (2009).
- Hansen, J., Autschbach, J. & Davies, H. M. L. Computational study on the selectivity of donor/acceptor-substituted rhodium carbenoids. *J. Org. Chem.* **74**, 6555–6563 (2009).

Supplementary Information is available in the online version of the paper.

**Acknowledgements** Financial support was provided by the National Science Foundation (NSF) through the CCI Center for Selective C–H Functionalization (CHE-1205646). We thank Novartis and AbbVie for supporting our research in C–H functionalization. We thank D. Guptill for conducting some preliminary studies on this project. D.G.M. gratefully acknowledges an NSF MRI-R2 grant (CHE-0958205) and the use of the resources of the Cherry Emerson Center for Scientific Computation.

**Author Contributions** K.L. performed and analysed the majority of the synthetic experiments. S.N. prepared the first meta-disubstituted catalyst, D.G.M. conducted the computational studies and J.B. conducted the X-ray crystallographic studies. K.L. and H.M.L.D. designed the synthetic experiments and prepared the manuscript.

**Author Information** The crystal data have been deposited in The Cambridge Crystallographic Data Centre (<http://www.ccdc.cam.ac.uk>) under accession number 1445448. Reprints and permissions information is available at [www.nature.com/reprints](http://www.nature.com/reprints). The authors declare competing financial interests: details are available in the online version of the paper. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to H.M.L.D. ([hmdavie@emory.edu](mailto:hmdavie@emory.edu)).

# Ancient micrometeorites suggestive of an oxygen-rich Archaean upper atmosphere

Andrew G. Tomkins<sup>1</sup>, Lara Bowlst<sup>1</sup>, Matthew Genge<sup>2,3</sup>, Siobhan A. Wilson<sup>1</sup>, Helen E. A. Brand<sup>4</sup> & Jeremy L. Wykes<sup>1,4,5</sup>

It is widely accepted that Earth's early atmosphere contained less than 0.001 per cent of the present-day atmospheric oxygen ( $O_2$ ) level, until the Great Oxidation Event resulted in a major rise in  $O_2$  concentration about 2.4 billion years ago<sup>1</sup>. There are multiple lines of evidence for low  $O_2$  concentrations on early Earth, but all previous observations relate to the composition of the lower atmosphere<sup>2</sup> in the Archaean era; to date no method has been developed to sample the Archaean upper atmosphere. We have extracted fossil micrometeorites from limestone sedimentary rock that had accumulated slowly 2.7 billion years ago before being preserved in Australia's Pilbara region. We propose that these micrometeorites formed when sand-sized particles entered Earth's atmosphere and melted at altitudes of about 75 to 90 kilometres (given an atmospheric density similar to that of today<sup>3</sup>). Here we show that the FeNi metal in the resulting cosmic spherules was oxidized while molten, and quench-crystallized to form spheres of interlocking dendritic crystals primarily of magnetite ( $Fe_3O_4$ ), with wüstite ( $FeO$ ) + metal preserved in a few particles. Our model of atmospheric micrometeorite oxidation suggests that Archaean upper-atmosphere oxygen concentrations may have been close to those of the present-day Earth, and that the ratio of oxygen to carbon monoxide was sufficiently high to prevent noticeable inhibition of oxidation by carbon monoxide. The anomalous sulfur isotope ( $\Delta^{33}S$ ) signature of pyrite ( $FeS_2$ ) in seafloor sediments from this period, which requires an anoxic surface environment<sup>4</sup>, implies that there may have been minimal mixing between the upper and lower atmosphere during the Archaean.

The low concentration of  $O_2$  in Earth's lower atmosphere during the Archaean period (3.9–2.5 Ga) has been demonstrated through (1) preservation of detrital pyrite and uraninite in ancient sediments derived from weathering of rocks on land (these weather rapidly in the presence of oxygen<sup>5</sup>), (2) a lack of oxidized iron in palaeosols<sup>6</sup>, and most definitively, (3) the strong mass-independent fractionation of sulfur isotopes recorded in seafloor pyrite<sup>4</sup>. In addition, large volumes of banded iron formations in late Archaean sedimentary rock sequences require high concentrations of dissolved  $Fe^{2+}$  to have existed in the Archaean oceans, which buffered atmospheric and oceanic  $O_2$  to very low concentrations<sup>7</sup>. After the evolution of photosynthesizing bacteria (possibly at ~2.7 Ga; refs 8, 9), this balance was maintained while there was sufficient dissolved  $Fe^{2+}$  in seawater, which was probably facilitated by particularly active volcanism in the period 2.7–2.4 Ga (ref. 7). The highly variable  $\Delta^{33}S$  signature of pyrite ( $FeS_2$ ) in carbonaceous shales from this period is evidence that volcanically released  $SO_2$  was dissociated by ultraviolet radiation in the atmosphere and then never re-oxidized; this requires the  $O_2$  concentration to have been <0.001% of the present atmospheric level<sup>4</sup>. The Great Oxidation Event is thought to have occurred when the rate of  $O_2$  production outstripped the rate of its removal via iron oxide sedimentation<sup>1</sup>, as a consequence of increased bacterial colonization and decreased volcanic activity<sup>7,10</sup>. After 2.4 Ga

the  $\Delta^{33}S$  of pyrite is invariant, indicating that atmospheric  $O_2$  had significantly increased<sup>11</sup>.

Sulfur has been injected continuously into Earth's atmosphere by volcanic eruptions over geological time. In the case of the largest ultraplinian eruptions, ejection columns can reach the upper stratosphere, rising to heights of 50 km (ref. 12). Today, volcanic material from a single eruption may remain in the stratosphere for years because there is little mixing within this layer<sup>13</sup>. On the other hand, micrometeorites entering the modern atmosphere experience their peak temperature at 75–90 km altitude<sup>14</sup> as they are decelerated from velocities exceeding  $12\text{ km s}^{-1}$  between 80 km and 150 km altitude<sup>15</sup>. For those micrometeorites that were sufficiently heated to melt completely, their high surface area results in quench crystallization as they cool at the end of deceleration. (We note that 70%–90% of modern micrometeorites with diameters >100  $\mu\text{m}$  are completely melted, whereas ~22% of those 25–50  $\mu\text{m}$  in size are fully or partially melted<sup>16</sup>.) Because small micrometeorites are melted and quench-crystallized within two seconds, modern micrometeorites only chemically interact with higher levels of the atmosphere. Therefore, we suggest that fossil micrometeorites trapped in sedimentary rocks represent a previously untapped, long-term record of the chemical composition of Earth's upper atmosphere.

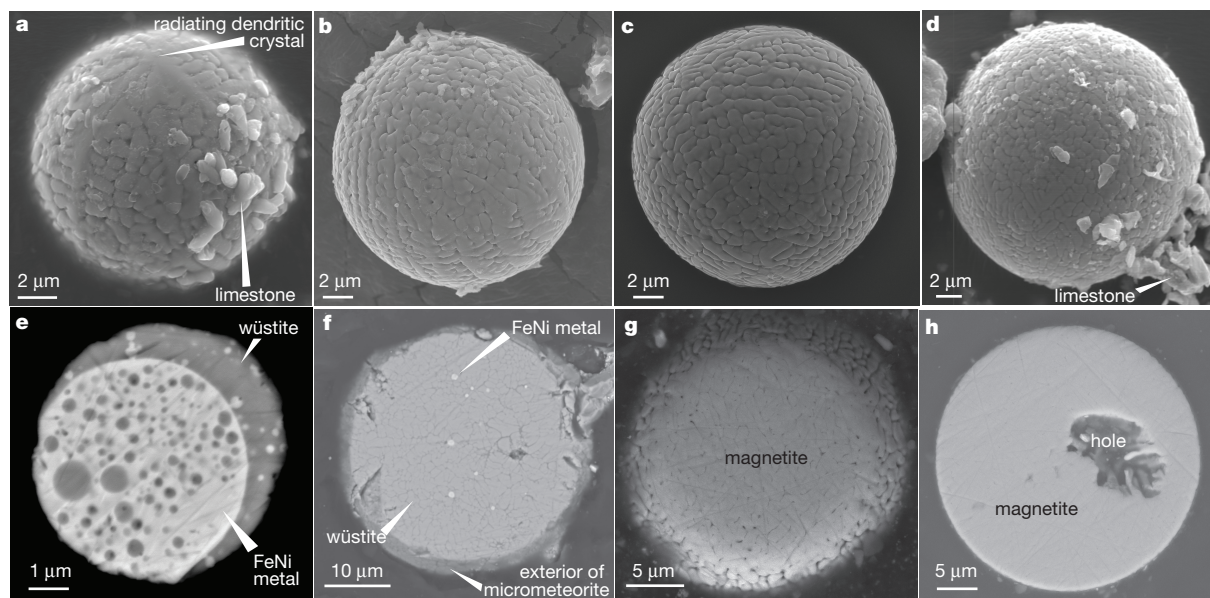
To obtain fossil micrometeorites that are older than the oldest so far found (which fell 1.8 Gyr ago<sup>17</sup>), we sampled limestone layers from the Meentheena Member from the Pilbara region of northwest Australia. The 30–50-m thick Meentheena Member is part of the Tumbiana Formation in the Mount Bruce Supergroup of the Hamersley Basin<sup>18</sup>; its age is  $19\,2721 \pm 4\text{ Ma}$ . We chose this unit because (1) it is barely affected by deformation and metamorphism, (2) its sediments were deposited slowly (a deposition rate of ~45 m  $\text{Myr}^{-1}$ ; ref. 20), thereby allowing greater accumulation of micrometeorites, and (3) limestone is easily dissolved with acid, allowing rapid extraction of micrometeorites from large samples. We preferentially selected samples from surface outcrops with fine horizontal laminations (from  $21^\circ 17' 43.8''\text{ S}$ ,  $120^\circ 27' 28.7''\text{ E}$ ; Extended Data Fig. 1) because these imply a slow deposition rate in a calm, deeper-water setting.

Sixty micrometeorites, 8.6–50  $\mu\text{m}$  in diameter, were separated from three limestone samples and examined by microanalytical techniques (Methods). All extracted micrometeorites are cosmic spherules — meteoritic material that was fully melted during atmospheric entry. Fifty-nine of these are I-type (iron) cosmic spherules (that is, they contain no silicate material, only FeNi metal and/or Fe oxide phases) and one is a 'glassy type' (that is, devitrified silicate glass). I-type cosmic spherules represent meteoritic FeNi metal that melted and reacted with the atmosphere during deceleration, via oxidation of Fe to form wüstite and magnetite<sup>16</sup>.

Figure 1 shows a representative selection of the recovered spherules; for comparison, examples of modern I-type spherules and weathered micrometeorites are shown in Extended Data Fig. 2. The exterior of all I-type spherules consists of interlocking dendritic crystals of magnetite

<sup>1</sup>School of Earth, Atmosphere and Environment, Monash University, Melbourne, Victoria 3800, Australia. <sup>2</sup>Impact and Astromaterials Research Centre, Department of Earth Science and Engineering, Imperial College London, Exhibition Road, London SW7 2AZ, UK. <sup>3</sup>Department of Mineralogy, The Natural History Museum, Cromwell Road, London SW7 2BT, UK. <sup>4</sup>Australian Synchrotron, 800 Blackburn Road, Clayton, Victoria 3168, Australia. <sup>5</sup>Department of Earth and Planetary Sciences, Macquarie University, North Ryde, New South Wales 2113, Australia.





**Figure 1 | Examples of fossil micrometeorites recovered in this study.** **a–d**, Representative examples of surface features of the micrometeorites, characterized by interlocking dendritic iron oxide crystals. **e–h**, Representative examples of the internal features of the micrometeorites. The majority of micrometeorites that we sectioned were identical to the

magnetite-only examples shown in **g** and **h**. The metal-dominated sample in **e** is the only one we observed; the spheres of wüstite in metal indicate mingling of immiscible melts. The small size of this example is consistent with the enhanced preservation of metal in smaller micrometeorites predicted by the oxidation modelling (Fig. 2).

( $\text{Fe}^{2+}\text{Fe}^{3+}_2\text{O}_4$ ) and occasionally, the less oxidized wüstite ( $\text{FeO}$ ). Some spherules have limestone still attached, indicating that they are unlikely to represent modern contamination. The interiors of sectioned I-type spherules vary from being magnetite-only (9 of 11 micrometeorites) to wüstite with FeNi metal (2 of 11 micrometeorites; Fig. 1, Extended Data Table 1, Extended Data Fig. 3), identified via wavelength dispersive X-ray spectroscopy (WDS) and Raman spectroscopy. Synchrotron powder X-ray diffraction of two additional micrometeorites found only magnetite throughout the entire volume of both micrometeorites (Extended Data Fig. 4).

Evidence that the recovered specimens are micrometeorites that were oxidized in the atmosphere, rather than products of diagenesis, metamorphism or later weathering includes: (1) the spheres of FeNi metal encased in wüstite (Fig. 1) confirm an extraterrestrial origin (FeNi metal does not form, and wüstite is rare, in sedimentary environments on Earth) and indicate quenching of a partially oxidized metallic liquid; (2) the spherical morphology and high surface area of the interlocking magnetite crystals indicate rapid crystallization from a liquid with high surface tension, these are typical of modern micrometeorites recovered from Antarctic ice<sup>16</sup> and the deep ocean<sup>21</sup>, and inconsistent with oxidation of iron metal encased in sediment<sup>22</sup>, where there is no mechanism to promote retention of the spherical shape (for example, Extended Data Fig. 2e); and (3) oxidation of metal during modern weathering produces haematite, ferrihydrite and goethite, and results in volume expansion sufficient to destroy the delicate submicrometre surface textures shown in Fig. 1.

Given their oxidized mineralogy, we designed a mathematical model to examine micrometeorite oxidation during atmospheric entry (see Methods). Molecules in the Archaean upper atmosphere capable of causing oxidation of Fe metal include  $\text{O}_2$  and  $\text{CO}_2$  (Extended Data Fig. 5), whereas CO is the only moderately abundant species capable of causing reduction; other species were likely to have been present at concentrations too low to be relevant<sup>23</sup>. Given that nearly all of the I-type cosmic spherules from the Meentheena Limestone contain dominant proportions of magnetite and/or wüstite relative to metal, the oxidizing molecules must have dominated Earth's upper atmosphere at 2.72 Ga. Equilibrium-based calculations imply that small micrometeorites cannot be oxidized to liquid magnetite by a  $\text{CO}_2$ -dominated atmosphere, and that some  $\text{O}_2$  is required (Extended Data Fig. 5). But the

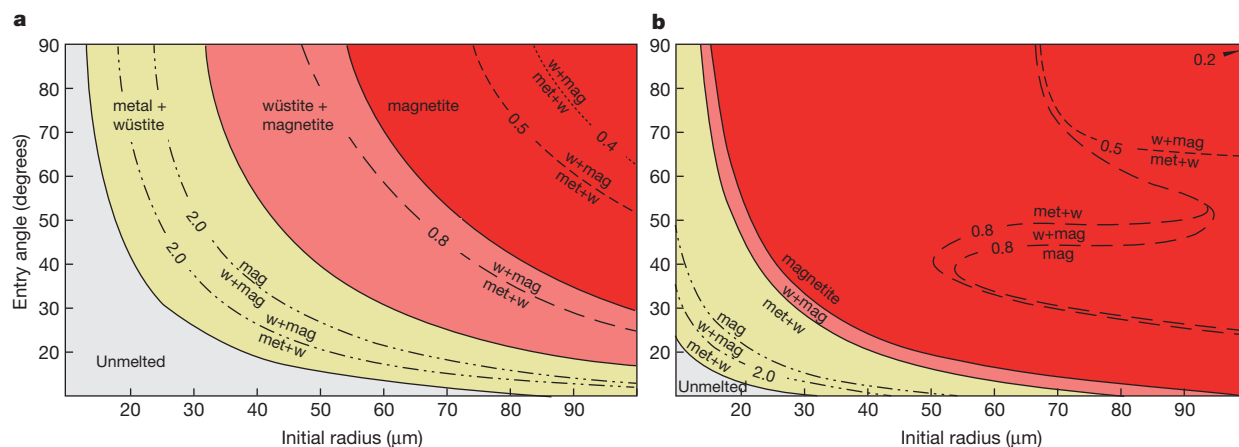
relative concentrations of  $\text{O}_2$  and  $\text{CO}_2$  are challenging to model, because the short duration of atmospheric heating means that equilibrium is never reached. Nonetheless, experiments on high temperature iron oxidation in mixed gases show that oxidation in the presence of  $\text{O}_2$  is rapid and readily forms magnetite + wüstite, whereas oxidation driven by  $\text{CO}_2$  alone is slow and produces only wüstite (Extended Data Fig. 6)<sup>24,25</sup>. The predominance of magnetite in the Archaean micrometeorites thus implies that  $\text{O}_2$  was an important oxidizing species. Furthermore, experiments indicate that when oxygen is present, addition of  $\text{CO}_2$  has little effect on brief oxidation events<sup>24</sup>. Our modelling therefore investigates how much oxygen would need to be present to generate the observed micrometeorite mineralogy.

Modelling of entry heating and oxidation of iron micrometeorites is complicated by competition between Fe oxidation and evaporative removal of the iron-oxide exterior; evaporation is greater for larger particles, faster entry velocities and steeper entry angles (Extended Data Fig. 7). Nonetheless, owing to the high temperatures, molten state and high surface area, reaction between metal and the atmosphere is rapid. Consequently, iron-oxide-free metal spherules have not been found amongst thousands of modern micrometeorites<sup>26</sup>. There are, however, modern examples of incomplete equilibration where a thick magnetite-wüstite shell surrounds a metal bead<sup>16</sup>.

Figure 2 shows the results of atmospheric oxidation modelling for I-type micrometeorites, and demonstrates that significant survival of metal occurs when atmospheric oxygen abundance is lower than 0.2 times the present atmospheric level (PAL), particularly in the case of smaller micrometeorites. In contrast, oxide- and particularly magnetite-dominated spherules are most abundant at atmospheric  $\text{O}_2$  concentrations similar to the Earth's current atmosphere. The absence of metal in the majority of the sectioned Archaean spherules, coupled with their small size, therefore indicates that they formed by heating in an atmosphere with a dramatically higher proportion of oxygen than the estimated surface concentration of oxygen at this time ( $\leq 1 \times 10^{-5}$  PAL<sup>4</sup>). Assuming that current models of the Archaean lower atmosphere are accurate, our results imply strong decoupling of the lower and upper atmosphere at that time.

In today's atmosphere there is rapid, vigorous mixing below the tropopause (altitude 8–18 km), little vertical mixing within the stratosphere (vertical diffusivity is  $\sim 0.1 \text{ m}^2 \text{ s}^{-1}$ ; ref. 27) below the stratopause





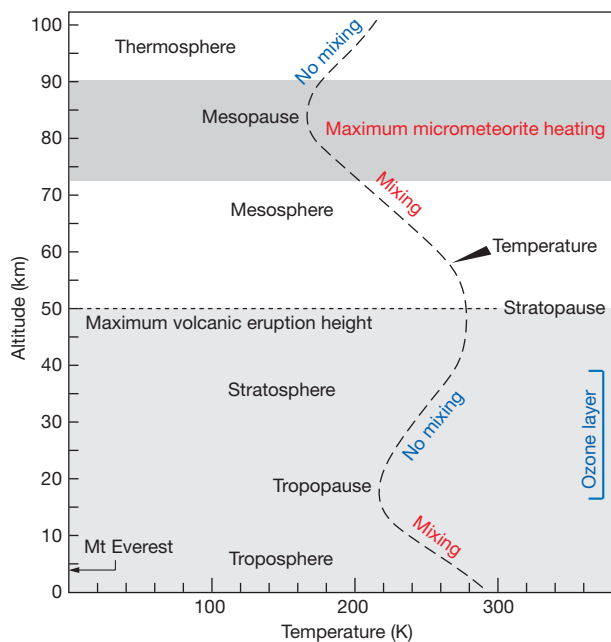
**Figure 2 | Results of the atmospheric oxidation model.** Shown in colour are the stability fields for the dominant iron-bearing phases after heating relative to the size and entry angle of particles. Abbreviations: met, metal; w, wüstite; mag, magnetite. **a, b**, Results are given for two different entry velocities;  $12 \text{ km s}^{-1}$  (**a**) is the minimum possible entry velocity,

and  $18 \text{ km s}^{-1}$  (**b**) is a common velocity for meteoroids that survive atmospheric entry. The coloured stability fields (olive, pink, red) are for the current atmospheric oxygen abundance; dashed lines indicate the shift in these fields at different oxygen abundances (0.2, 0.5, 0.8, and 2.0 times PAL, numbers on curves).

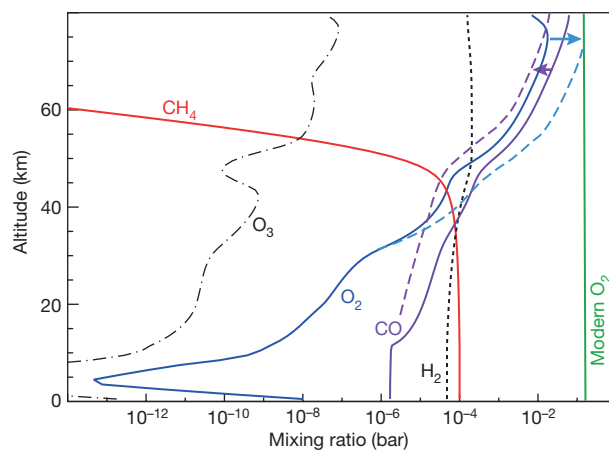
(50–55 km), and moderate vertical mixing within the mesosphere up to the mesopause (85 km), above which there is minimal mixing again<sup>13</sup>. The extent of vertical atmospheric mixing is controlled by the thermal properties of the gases; because hotter air is less dense than colder air, mixing is inhibited where air temperature increases with altitude, as in the stratosphere where more ultraviolet radiation is absorbed and converted to heat by ozone at higher altitudes (Fig. 3). Importantly, although mixing occurs within the mesosphere, there is minimal chemical communication between the altitude of maximum I-type micrometeorite heating (75–90 km) and that of volcanic flux (in the troposphere and stratosphere, 0–50 km)<sup>13</sup>, largely due to the inverted temperature profile of the stratosphere. Our observations suggest that a similar atmospheric structure with inhibited vertical mixing existed at 2.72 Ga. Kasting<sup>23</sup> suggested that biosphere–atmosphere interactions may have moderated the  $\text{CH}_4:\text{CO}_2$  ratio to stabilize an Archaean atmosphere with an optically thin organic haze; we suggest that if such

a haze existed it would absorb heat and create an inverted temperature profile (see, for example, ref. 28) thereby suppressing mixing between the oxygen-poor lower and oxygen-rich upper atmosphere.

Some chemical models of the Archaean atmosphere suggest that a methane-bearing atmosphere could have existed up to ~50 km altitude (Fig. 4)<sup>29</sup>, and this could represent the hypothesized transition between poorly mixing upper and lower atmospheric domains. However, micrometeorites cannot test the altitude of this transition since they are quench-crystallized by the time they reach ~75 km. The only work similar to that presented here is an analysis of the oxidation state of impact ejecta preserved in 3.24-Gyr-old spherule beds, which found considerable heterogeneity in oxidation state and that atmospheric oxygen fugacity was  $< 10^{-4}$  bar at that time<sup>30</sup>. Although Krull-Davatzes *et al.* suggested that the oxidation state of impact ejecta is set within the ejection plume (see ref. 30 and references therein), and thus impact spherules represent samples of the lower atmosphere, it is plausible that some are remelted during re-entry<sup>31</sup>. Because impact spherules are up to two orders of magnitude larger than micrometeorites and



**Figure 3 | Profile of the modern atmosphere.** Shown are the altitudes reached by volcanic eruptions (lower grey shaded area) compared with the altitudes of maximum micrometeorite heating (upper grey shaded area), as well as the temperature profile (dashed curve) and how this influences vertical mixing.



**Figure 4 | A recent model of the Archaean atmosphere that includes methane, showing the effects of  $\text{CO}_2$  photolysis on  $\text{O}_2$  and CO concentration.** The solid blue and purple lines indicate the estimated  $\text{O}_2$  and CO concentrations of the existing model<sup>29</sup>, whereas the dashed blue and purple lines (arrowed) approximate the relatively small shifts in the positions of these lines needed to satisfy the observed micrometeorite oxidation. Modern  $\text{O}_2$  (green line) is shown for reference;  $\text{H}_2$  and  $\text{O}_3$  concentrations are too low to affect micrometeorite chemistry. This figure is modified from ref. 29 with permission from John Wiley and Sons, license number 3816161033966.

re-enter the atmosphere at lower velocity, those that remelted would sample a thicker belt and lower level of the atmosphere, possibly close to the hypothesized oxygen transition. Detailed modelling of impact spherule re-entry thus represents a possible avenue for sampling the chemistry of middle levels of the atmosphere; in particular, larger particles are more likely to have been remelted, so certain size fractions may have sampled appropriate atmospheric levels.

Chemical models of the Archaean atmosphere are based on gas fluxes from modern volcanoes and the effects of ultraviolet photolysis on those gases, and constrained by the above observations from ancient rocks<sup>23</sup>. These models imply that at lower atmospheric levels, reaction with abundant volcano-derived H<sub>2</sub> gas eliminated free O<sub>2</sub>, consistent with the  $\Delta^{33}\text{S}$  data that require an oxygen-poor lower atmosphere. Above an altitude of ~50 km, photolysis of CO<sub>2</sub> produced O<sub>2</sub> and CO (Fig. 4; ref. 29), but there is extensive debate about how much CO<sub>2</sub> was present in the Archaean atmosphere<sup>23</sup>, with estimates ranging from 10 to 1,000 PAL. The more CO<sub>2</sub> present in these models, the more O<sub>2</sub> and CO is produced by photolysis in the upper atmosphere. Furthermore, these models suggest that there may have been more CO produced than O<sub>2</sub>; for example, the model in Fig. 4 suggests an upper atmosphere with an O<sub>2</sub>/CO ratio of ~0.2. However, this abundant CO is inconsistent with the observed magnetite-rich micrometeorites because it would act as a reductant (for example, CO + FeO = CO<sub>2</sub> + Fe; Extended Data Fig. 5). At equilibrium, the point where oxidation balances reduction comes when there is half as much O<sub>2</sub> as CO, so any oxidation implies O<sub>2</sub>/CO > 0.5. If the O<sub>2</sub>/CO ratio was only slightly greater than 0.5, the oxidation would have been sluggish and we would only see thin wüstite rims on metal in the larger spherules. Complete oxidation of the very small micrometeorites to magnetite suggests that the Archaean upper atmosphere may have been characterized by O<sub>2</sub> levels approaching those of the present-day Earth, requiring relatively abundant CO<sub>2</sub> as a source of O<sub>2</sub>, and elevated ratios of O<sub>2</sub> to CO.

**Online Content** Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

**Received 30 May 2015; accepted 9 March 2016.**

- Lyons, T. W., Reinhard, C. T. & Planavsky, N. J. The rise of oxygen in Earth's early ocean and atmosphere. *Nature* **506**, 307–315 (2014).
- Farquhar, J., Zerkle, A. L. & Bekker, A. in *Treatise in Geochemistry* Vol. 6, *The Atmosphere – History* (eds Holland, H. & Turekian, K.) 91–138 (2014).
- Som, S. M., Catling, D. C., Harnmeijer, J. P., Polivka, P. M. & Buick, R. Air density 2.7 billion years ago limited to less than twice modern levels by fossil raindrop imprints. *Nature* **484**, 359–362 (2012).
- Pavlov, A. A. & Kasting, J. F. Mass-independent fractionation of sulfur isotopes in Archean sediments: strong evidence for an anoxic Archean atmosphere. *Astrobiology* **2**, 27–41 (2002).
- Rasmussen, B. & Buick, R. Redox state of the Archean atmosphere: evidence from detrital heavy minerals in ca. 3250–2750 Ma sandstones from the Pilbara Craton, Australia. *Geology* **27**, 115–118 (1999).
- Rye, R. & Holland, H. D. Paleosols and the evolution of atmospheric oxygen: a critical review. *Am. J. Sci.* **298**, 621–672 (1998).
- Bekker, A. *et al.* Iron formation: the sedimentary product of a complex interplay among mantle, tectonic, and biospheric processes. *Econ. Geol.* **105**, 467–508 (2010).
- Buick, R. The antiquity of oxygenic photosynthesis: evidence from stromatolites in sulphate-deficient Archean lakes. *Science* **255**, 74–77 (1992).
- Eigenbrode, J. L. & Freeman, K. H. Late Archean rise of aerobic microbial ecosystems. *Proc. Natl Acad. Sci. USA* **103**, 15759–15764 (2006).
- Konhauser, K. O. *et al.* Oceanic nickel depletion and a methanogen famine before the Great Oxidation Event. *Nature* **458**, 750–753 (2009).
- Farquhar, J. & Wing, B. A. Multiple sulfur isotopes and the evolution of the atmosphere. *Earth Planet. Sci. Lett.* **213**, 1–13 (2003).
- Cioni, R., Marianelli, P., Santacroce, R. & Sbrana, A. in *Encyclopedia of Volcanoes* (ed. Sigurdsson, H.) 477–494 (Academic, 2000).
- Finlayson-Pitts, B. & Pitts, J. *Chemistry of the Upper and Lower Atmosphere: Theory, Experiments, and Applications* (Elsevier, 1999).
- Love, S. G. & Brownlee, D. E. Heating and thermal transformation of micrometeoroids entering the Earth's atmosphere. *Icarus* **89**, 26–43 (1991).
- Rietmeijer, F. J. M. & Nuth, J. A. Collected extraterrestrial materials: constraints on meteor and fireball compositions. *Earth Moon Planets* **82–83**, 325–350 (2000).
- Genge, M., Engrand, C., Gounelle, M. & Taylor, S. The classification of micrometeorites. *Meteorit. Planet. Sci.* **43**, 497–515 (2008).
- Tianrui, S., Zhengjun, H., Yusheng, W. & Yanxue, L. A study of Mesoproterozoic iron cosmic micro-spherules from 1.8 Ga and 1.6 Ga old strata in the Ming Tombs District, Beijing. *Acta Geol. Sin.* **81**, 649–657 (2007).
- Awramik, S. M. & Buchheim, H. P. A giant, Late Archean lake system: the Meentheena Member (Tumbiana Formation; Fortescue Group), Western Australia. *Precamb. Res.* **174**, 215–240 (2009).
- Blake, T. S., Buick, R., Brown, S. J. A. & Barley, M. E. Geochronology of a Late Archaean flood basalt province in the Pilbara Craton, Australia: constraints on basin evolution, volcanic and sedimentary accumulation, and continental drift rates. *Precamb. Res.* **133**, 143–173 (2004).
- Trendall, A. F., Compston, W., Nelson, D. R., De Laeter, J. R. & Bennett, V. C. SHRIMP zircon ages constraining the depositional chronology of the Hamersley Group, Western Australia. *Aust. J. Earth Sci.* **51**, 621–644 (2004).
- Rudraswami, N. G. *et al.* Refractory metal nuggets in different types of cosmic spherules. *Geochim. Cosmochim. Acta* **131**, 247–266 (2014).
- Chevrier, V., Rochette, P., Mathe, P.-E. & Grauby, O. Weathering of iron-rich phases in simulated Martian atmospheres. *Geology* **32**, 1033–1036 (2004).
- Kasting, J. F. in *Treatise in Geochemistry* Vol. 6, *The Atmosphere – History* (eds Holland, H. & Turekian, K.) 157–175 (2014).
- Abuluwafa, H. T., Guthrie, R. I. L. & Ajersch, F. Oxidation of low carbon steel in multicomponent gases: Part I. Reaction mechanisms during isothermal oxidation. *Metall. Mater. Trans A* **28**, 1633–1641 (1997).
- Bredesen, R. & Kofstad, P. On the oxidation of iron in CO<sub>2</sub> + CO mixtures. III: Coupled linear parabolic kinetics. *Oxidat. Metals* **36**, 27–56 (1991).
- Genge, M. The origins of I-type spherules and the atmospheric entry of iron micrometeoroids. *Meteorit. Planet. Sci.* (in the press).
- Legras, B., Joseph, B. & Lefevre, F. Vertical diffusivity in the lower stratosphere from Lagrangian back-trajectory reconstructions of ozone profiles. *J. Geophys. Res.* **108**, 4562, <http://dx.doi.org/10.1029/2002JD003045> (2003).
- Fulchignoni, M. *et al.* In situ measurements of the physical characteristics of Titan's environment. *Nature* **438**, 785–791 (2005).
- Zahnle, K., Claire, M. & Catling, D. The loss of mass-independent fractionation in sulfur due to a Palaeoproterozoic collapse of atmospheric methane. *Geobiology* **4**, 271–283 (2006).
- Krull-Davatzes, A. E., Byerley, G. R. & Lowe, D. R. Evidence for a low-O<sub>2</sub> Archean atmosphere from nickel-rich chrome spinels in 3.24 Ga impact spherules, Barberton greenstone belt, South Africa. *Earth Planet. Sci. Lett.* **296**, 319–328 (2010).
- Goldin, T. J. & Melosh, H. J. Self-shielding of thermal radiation by Chicxulub impact ejecta: firestorm or fizzle? *Geology* **37**, 1135–1138 (2009).

**Acknowledgements** We thank N. Wilson and A. Langendam for assistance with electron microprobe work and electron microscopy, respectively. The authors acknowledge use of the Monash Centre for Electron Microscopy, and use of the CSIRO Microbeam Laboratory. Part of this research was undertaken on the powder diffraction beamline at the Australian Synchrotron, Victoria, Australia. M.G. acknowledges STFC grant number ST/J001260/1.

**Author Contributions** A.G.T. conceptualized the project, conducted fieldwork and EMP analysis, and wrote the paper. L.B. conducted fieldwork, micrometeorite separation and SEM analysis. M.G. generated the micrometeorite oxidation model. S.A.W. advised on micrometeorite separation, conducted Raman spectroscopy and interpreted synchrotron results. H.E.A.B. conducted the synchrotron analysis. J.L.W. modelled the oxidizing conditions imposed at equilibrium by different atmospheres. All authors reviewed the paper prior to submission.

**Author Information** Reprints and permissions information is available at [www.nature.com/reprints](http://www.nature.com/reprints). The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to A.G.T. ([andrew.tomkins@monash.edu](mailto:andrew.tomkins@monash.edu)).

## METHODS

**Extraction and characterization.** The weathered exterior of the limestone blocks was removed using a diamond saw. The resulting samples (3.6 kg) were then crushed to <1 cm fragments, then bathed in 10% HCl (samples 1.1, 1.3a) or 20% HCl (sample 1.2) for 1–2 days. After wet sieving the residues into 20–125, 125–355, 355–1,000 and >1,000  $\mu\text{m}$  size fractions, magnetic separation and hand picking under a binocular microscope were used to isolate micrometeorites. Care was taken at all stages to keep the samples covered, to avoid contamination by present day micrometeorites. The isolated grains were then confirmed or rejected as micrometeorites through imaging using a JEOL 7001F FEG-SEM. Micrometeorites were then embedded in epoxy and sectioned for further imaging and electron microprobe (EMP) analysis using a JEOL 8500F HyperProbe operating at 15 kV accelerating voltage and 20 nA beam current. Eleven micrometeorites were sectioned effectively due to their small size.

Raman spectroscopy was used to confirm the EMP-determined identity of wüstite in Fig. 1f. Raman spectra were collected using a Renishaw 'Invia' Raman microscope fitted with a coherent 632.8 nm HeNe laser. The  $520.5\text{ cm}^{-1}$  band of a silicon wafer was used to calibrate the instrument before analysis of samples. Spectra were recorded between 2,000 and  $100\text{ cm}^{-1}$  with a resolution of  $1\text{--}2\text{ cm}^{-1}$  using a  $50\times$  objective lens. An initial spectrum was recorded using a power of  $\sim 1.5\text{ mW}$  at the sample surface and an exposure time of 60 s for 4 accumulations. The resulting spectrum (Extended Data Fig. 3a) is characterized by a broad, asymmetrical band at  $650\text{--}660\text{ cm}^{-1}$ , which is consistent with the spectrum of wüstite<sup>32</sup> (Extended Data Fig. 3b). Although magnetite also produces a band in the vicinity of  $660\text{--}680\text{ cm}^{-1}$ , the diagnostic peaks of magnetite at  $\sim 300$  and  $\sim 535\text{ cm}^{-1}$  are not observed in this spectrum. A second test of the wüstite identification is possible because wüstite decomposes rapidly to haematite when exposed to 632.8 nm radiation at  $\sim 7.0\text{ mW}$ , whereas magnetite is stable at these conditions<sup>32</sup>. Thus a second spectrum was recorded using  $\sim 7.5\text{ mW}$  at the sample surface, and a single accumulation with an exposure time of 50 s. In this acquisition, the targeted iron oxide phase immediately decomposed to produce the characteristic spectrum of haematite (Extended Data Fig. 3a), confirming the identity of wüstite.

An additional six whole micrometeorites were mounted on Hampton Scientific CryoLoops to investigate whether multiple phases could be identified within whole micrometeorite samples via powder X-ray diffraction (XRD). Data were collected on these using the Powder Diffraction beamline at the Australian Synchrotron. The penetration depth of the X-ray beam employed (15.0 keV) is on the order of  $50\text{ }\mu\text{m}$  in magnetite, which allows for non-destructive sampling of the entire volume of individual micrometeorites. Data were collected with the MYTHEN-II strip detector. Mineral phases were identified with reference to the ICDD PDF-2 database using the EVA V.1 software package available from Bruker AXS. Owing to their small size, reliable XRD data could only be obtained for two micrometeorites; results are shown in Extended Data Fig. 4).

To evaluate the oxidizing potential of different atmospheres, equilibrium gas speciation (and hence oxygen fugacity,  $f\text{O}_2$ ) was determined by Gibbs free energy minimization calculations using the HSC Chemistry (v. 6.1) software package. Modelling was undertaken between 1,250 and  $2,500^\circ\text{C}$ , and  $1$  to  $1 \times 10^{-6}$  bar pressure, assuming ideal gas mixing. For the  $\text{CO}_2\text{--CO}$  and  $\text{CO}_2$  gas mixtures species were limited to  $\text{O}_2$ , CO and  $\text{CO}_2$ , and  $\text{N}_2$  is considered to have no oxidizing or reducing capacity. Air species included in the modelling were CO,  $\text{CO}_2$ , H,  $\text{H}_2$ , HNO,  $\text{HNO}_2$ ,  $\text{HNO}_3$ ,  $\text{HO}_2$ ,  $\text{H}_2\text{O}$ , N,  $\text{N}_2$ , NO,  $\text{NO}_2$ ,  $\text{NO}_3$ ,  $\text{N}_2\text{O}$ , O,  $\text{O}_2$ ,  $\text{O}_3$ , OH, Ar. Results for the range of dynamic ram pressures experienced by micrometeorites are shown in Extended Data Fig. 5.

**Modelling.** The model used to calculate the extent of oxidation of iron metal during atmospheric entry is based on the model of ref. 14 and numerically integrates the partial differential equations for motion, heat and mass of particles decelerating in the atmosphere. Evaporation of particles is modelled using the Langmuir equation with coefficients for iron metal and iron oxide melts derived from experimental studies<sup>33</sup>. Evaporative mass loss is used in the model to calculate the change in mass and radius of the particle, which directly influences its deceleration. The average density of the particles is assumed to be that of wüstite, given that most of the observed particles are dominated by iron oxides and that the density of wüstite

is between that of magnetite and iron. The latent heat of evaporation is considered in evaluating the temperature of the particle, together with heat loss by thermal radiation and energy input due to incident atmospheric molecules. As in Love and Brownlee<sup>14</sup>, the trajectory of the particle, derived from the equations of motion, is used to trace change in altitude at each time step and thus atmospheric density. Time steps of 0.01 s were used for these simulations; results are independent of time step size to within 1% for the peak temperature of particles. The degree of oxidation is calculated from the mass of oxygen that particles encounter during deceleration, moderated by loss of oxygen by evaporation of iron oxide liquid. This calculation recognizes that the amount of oxygen accreted by iron micrometeorites can be no more than the mass of oxygen the particles encounter during deceleration. It also uses the observation that iron oxide liquid mantles the iron metal core of the particle due to surface tension effects resulting in evaporation of oxide rather than metal. Accretion of oxygen is also assumed to only occur when particles are molten since diffusion rates within solid metal are too slow to allow noticeable oxidation over the few seconds of deceleration.

Given the dependence of the extent of oxidation on both particle temperature and mass of oxygen encountered, the atmospheric density–altitude profile is important. In the absence of a consensus density model of the Archaean atmosphere, the 1976 US Standard Atmosphere is used in the simulations. Some models of the Archaean atmosphere, however, suggest that density variation with altitude was similar to the modern day<sup>3</sup>. Changes in the exact atmospheric density with altitude, however, can be seen to change the mass of gas encountered over molten flight relatively little. A denser Archaean atmosphere results in a larger mass of gas encountered per second, but particle deceleration would also occur more rapidly, giving a shorter period for accretion of oxygen. Conversely a less dense Archaean atmosphere results in a smaller mass of gas encountered per second but less deceleration, resulting in a longer period over which particles accrete oxygen. The results of 12,000 simulations with and without oxidation suggest that iron micrometeoroids undergo maximum heating at altitudes equivalent to 65–90 km; that is, within the current day mesosphere.

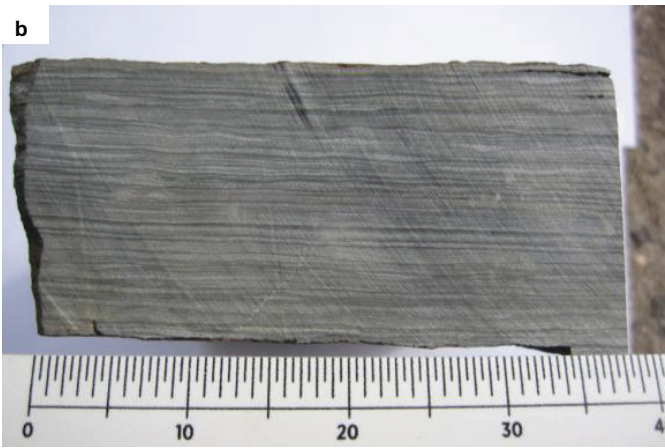
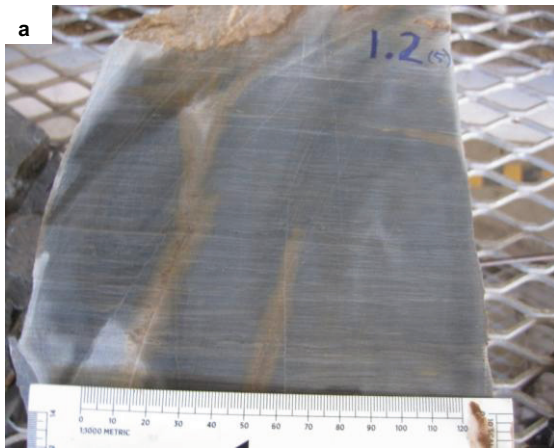
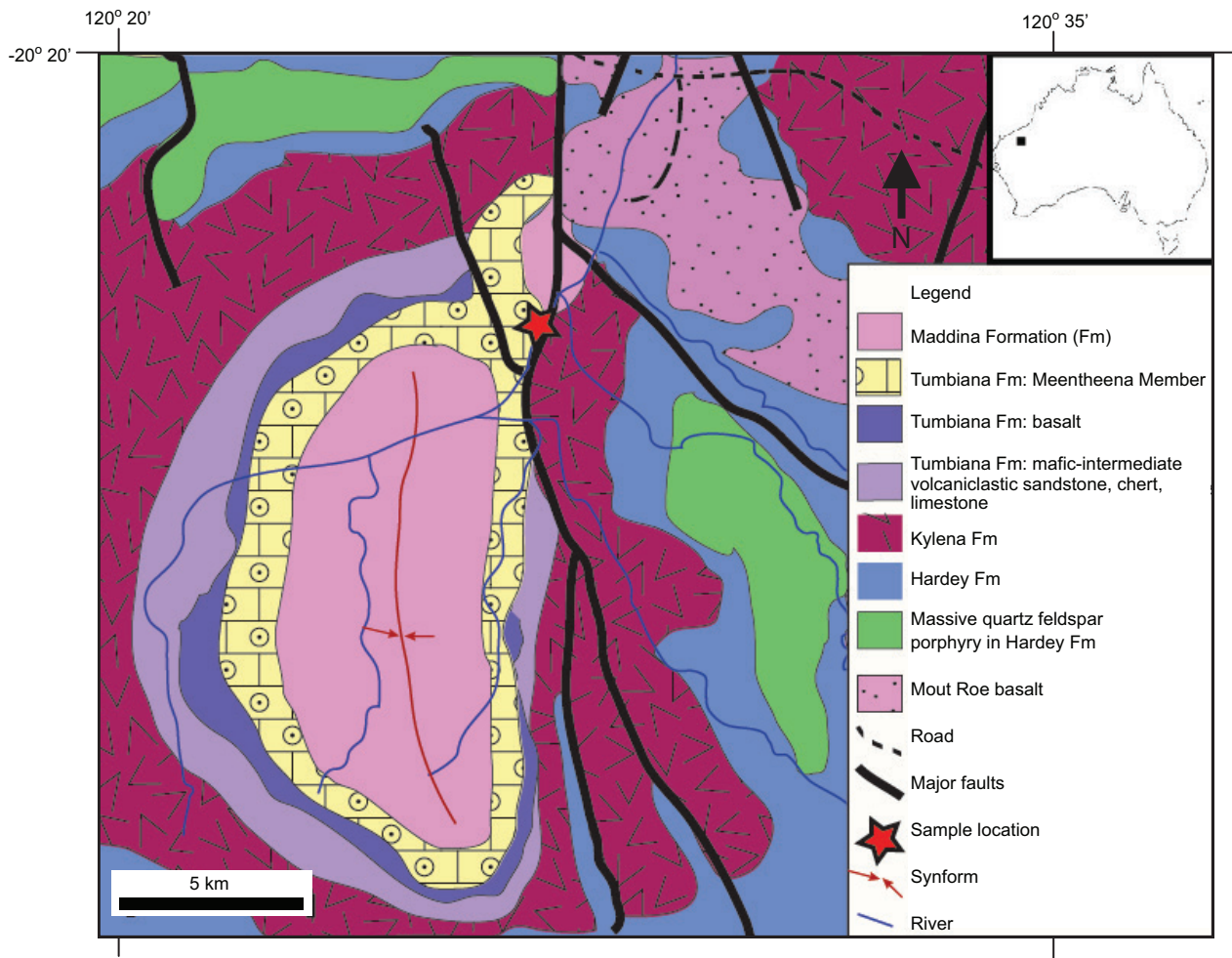
The abundance of oxygen in the Archaean atmosphere in the simulations was varied by scaling the modern day oxygen abundance with altitude. The effects of other gas species on oxidation or reduction were not considered within the model. Carbon monoxide, if present within the Archaean atmosphere, would oppose oxidation by driving reduction of iron oxide melt and production of  $\text{CO}_2$ . At equilibrium, if the  $\text{O}_2/\text{CO}$  ratio was  $<0.5$ , the metal would not be oxidized (that is,  $2\text{Fe} + \text{O}_2 = 2\text{FeO}$  versus  $\text{FeO} + \text{CO} = \text{Fe} + \text{CO}_2$ ). The abundances of oxygen used within the model can therefore be considered to be those in excess of reducing species in the atmosphere, although in a disequilibrium scenario of micrometeorite entry the rate of oxidation by  $\text{O}_2$  may be faster than the rate of reduction by CO. The important criteria within the oxidation model are the abundance of oxygen compared to that in the present atmosphere, the atmospheric density profile and the molten flight time of the particle, and the entry parameters that determine peak temperature and altitude of peak temperature. The objective of the simulations is not to provide an accurate and realistic model of Archaean atmospheric chemistry, but simply to estimate the abundance of free oxygen available to produce net oxidation, which is required to explain the presence of magnetite-rich Archaean I-type spherules. Even if the mass of oxygen accreted by particles predicted by the simulations is subject to an error of an order of magnitude, the results would still necessitate abundant oxygen within the Archaean upper atmosphere relative to the model shown in Fig. 4.

**Code availability.** The custom entry heating model code used in this paper was developed by M.G. and is not currently publicly available. For further details on the code see ref. 26.

**Sample size.** No statistical methods were used to predetermine sample size.

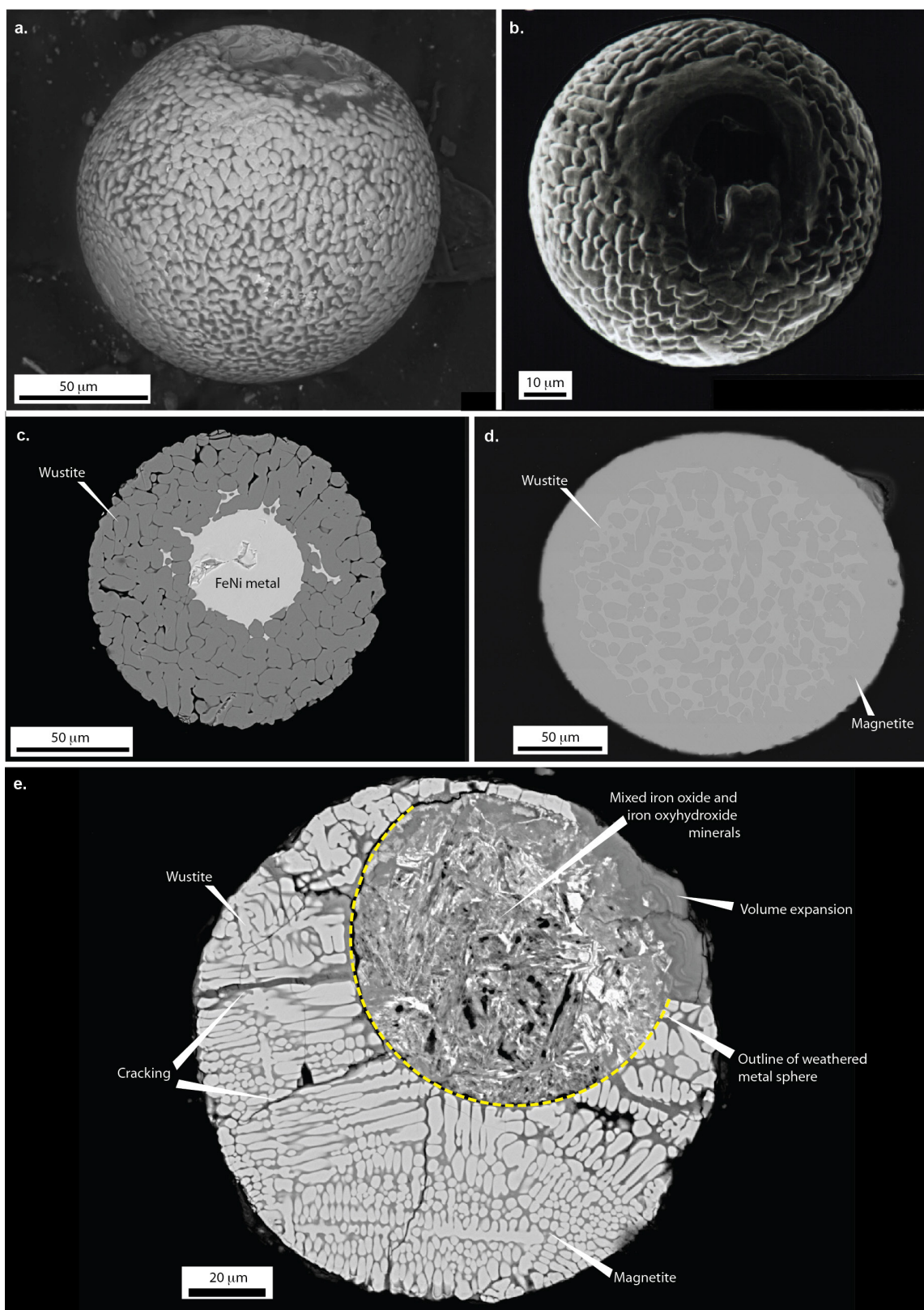
32. de Faria, D. L. A., Silva, S. V. & De Oliveira, M. T. Raman microspectroscopy of some iron oxides and oxyhydroxides. *J. Raman Spectrosc.* **28**, 873–878 (1997).
33. Wang, J., Davis, A. M., Clayton, R. N. & Mayeda, T. K. Kinetic isotopic fractionation during the evaporation of the iron oxide from liquid state. *Proc. Lunar Planet. Sci. Conf.* **25**, 1459–1460 (1994).





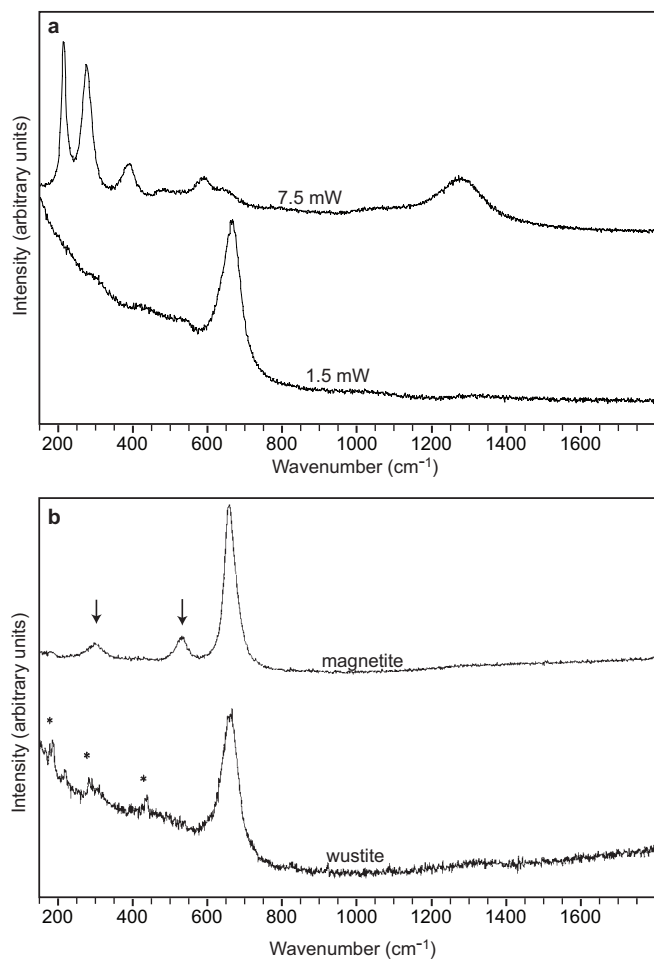
**Extended Data Figure 1 | Geological map showing the context of the sampling location (red star), and examples of the samples used in the study.** The samples in photos **a** and **b** show examples of the fine laminations that occur in some layers of this unit. In **a**, buff coloured zones along cracks highlight examples of modern day oxidative weathering, whereas the grey colouration of the remainder of the sample indicates that

it was not weathered, allowing survival of the micrometeorites. Weathered rock was removed using a diamond saw before micrometeorite separation. The geological map (top) is provided courtesy of the Geological Survey of Western Australia, Department of Mines and Petroleum. Copyright State of Western Australia, 2016.



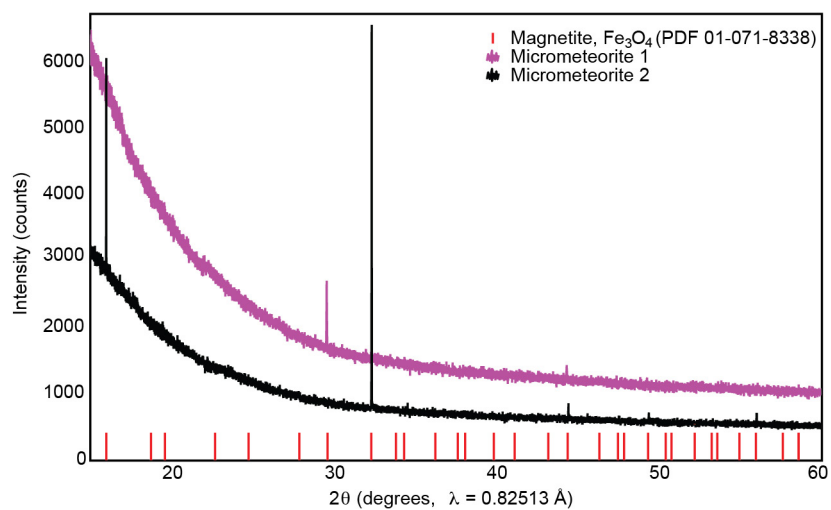
**Extended Data Figure 2 | Examples of modern iron-type micrometeorites collected from the Antarctic ice sheet. a, b,** Secondary electron images showing the typical exterior morphology. **c, d,** Back-scattered electron (BSE) images of polished cross-sections of two other micrometeorites, highlighting the interior mineralogical variation. **e,** BSE image showing an example of a partially weathered modern

micrometeorite in cross-section where the metal (outlined by the dashed yellow line) has been replaced by iron oxides and iron oxyhydroxides after arriving on the surface; expansion has led to cracking in the surrounding wüstite and magnetite, which would destroy the micrometeorite if continued. Note that the wüstite and magnetite are unaffected by the weathering. All imaging by M.G.



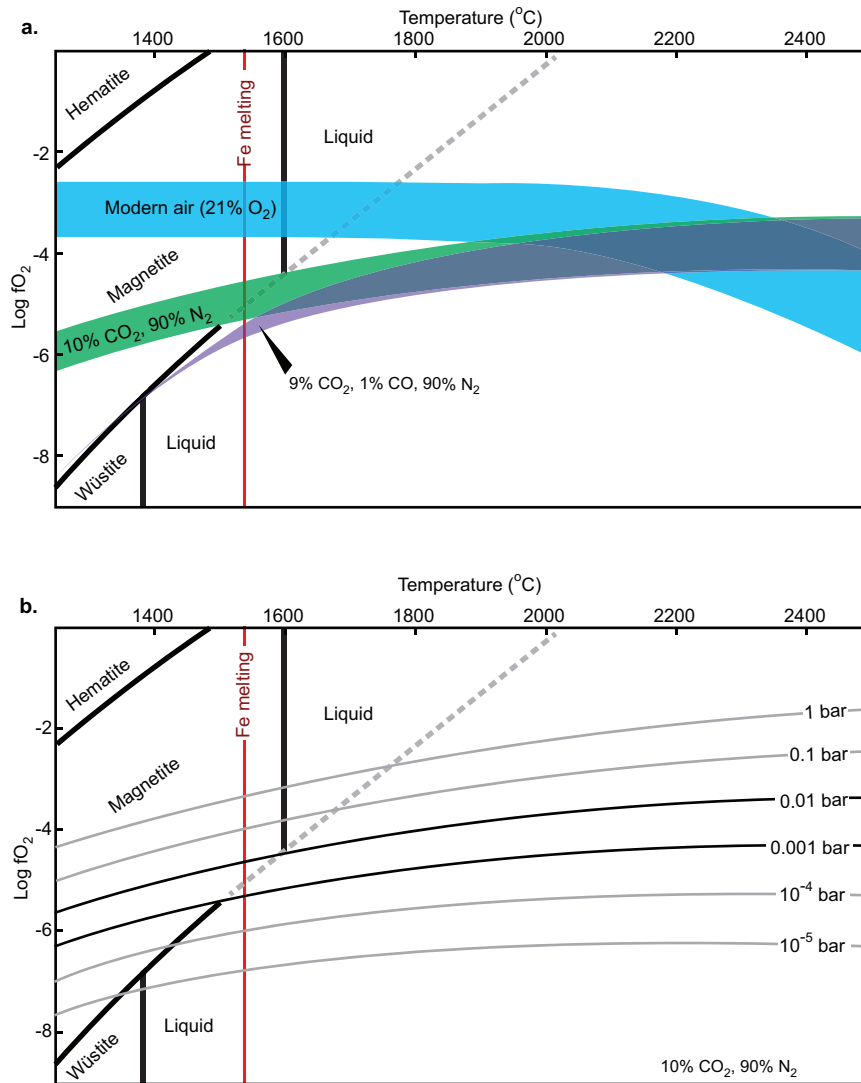
**Extended Data Figure 3 | Results of laser Raman spectroscopy confirming the identity of wüstite.** Wüstite is metastable below 570 °C and decomposes to haematite when higher laser power is used, whereas magnetite does not decompose in this fashion<sup>32</sup>. The iron oxide in the sectioned micrometeorite shown in Fig. 1f gave the spectrum shown in **a** at 1.5 mW and decomposed to produce the spectrum shown in **a** at 7.5 mW laser power, consistent with the characteristics of wüstite. Panel **b** shows the characteristic spectra of magnetite and wüstite from ref. 32; note the arrowed bumps that characterize magnetite, which are missing in wüstite, and the broadened main peak of wüstite. Panel **b** is modified from ref. 32 with permission from John Wiley and Sons, license number 3850481150510.





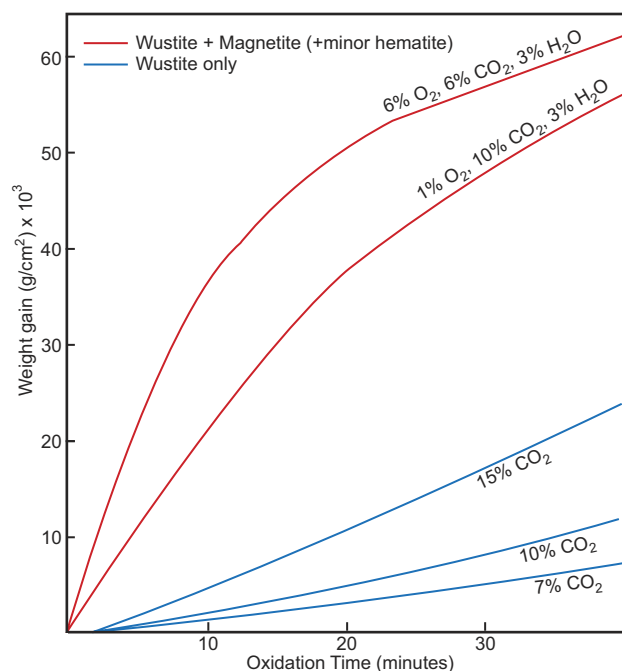
**Extended Data Figure 4 | Two synchrotron powder X-ray diffraction patterns, each collected from a single micrometeorite, and reference positions of magnetite peaks.** Not all major Bragg peaks for magnetite are detected, and deviation from the expected relative intensities of peaks

is observed, as a consequence of poor particle size statistics (PDF 01-071-8338 refers to a Powder Diffraction File from the International Centre for Diffraction Data database; <http://www.icdd.com/>).



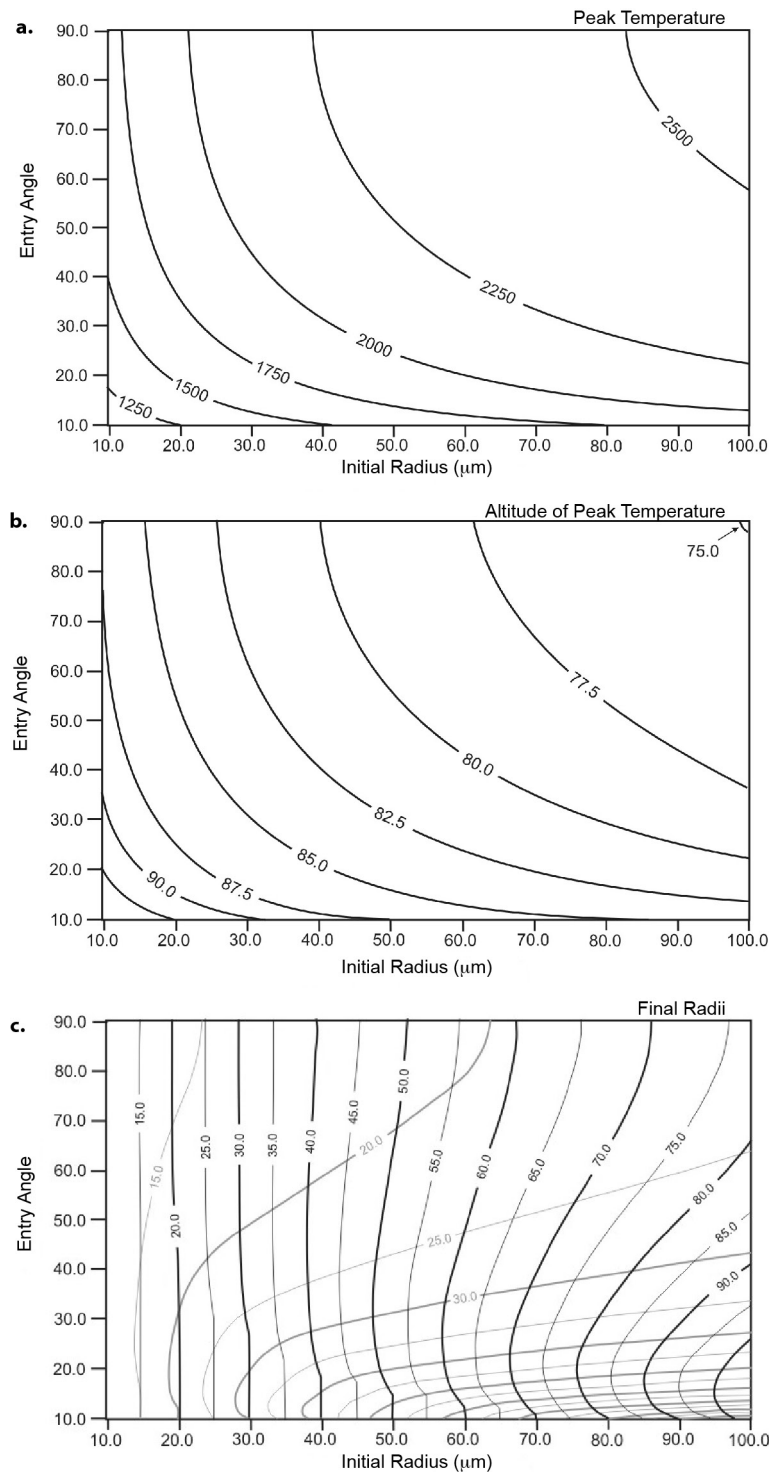
**Extended Data Figure 5 | Equilibrium modelling of the oxidizing conditions imposed by different atmospheres relative to the stability fields of haematite, magnetite and wüstite.** These models (oxygen fugacity versus temperature) represent the stability fields at equilibrium and do not consider the time needed to attain equilibrium (refer to Extended Data Fig. 6). In **a**, the top of each coloured band represents the conditions imposed by the maximum dynamic ram pressure (0.02 bar)

experienced by micrometeorites, which applies to the largest and fastest micrometeorites; the bottom of each band is more relevant to the small micrometeorites observed in this study (0.001 bar). In **b**, the model for an atmosphere containing 90%  $\text{N}_2$  and 10%  $\text{CO}_2$  is shown, with contours for pressure to allow a comparison with atmospheres of different  $\text{CO}_2$  abundance.



**Extended Data Figure 6 | Results of experiments on oxidation of low carbon steel in different gas mixes.** These are the results of experiments conducted at 1,100 °C and atmospheric pressure (1 bar); the gas mix is shown on the curves, the remaining gas being  $\text{N}_2$ . Oxidation was measured by progressive weight gain over time. The conditions of survivable micrometeorite entry are 400–2,800 °C and dynamic ram pressure of the order of 0.001–0.02 bar for <2 s. Higher temperatures result in more effective oxidation, lower pressures result in less effective oxidation. This figure was compiled by amalgamating figures 1 and 8 from ref. 24, with permission from Springer, license number 3850591244373.





#### Extended Data Figure 7 | Results of atmospheric entry modelling.

**a–c,** Plots of entry angle in degrees versus initial radius of micrometeorite in  $\mu\text{m}$ . **a,** Peak temperatures reached by I-type spherules (numbers on curves in  $^{\circ}\text{C}$ ) entering the atmosphere at  $12 \text{ km s}^{-1}$ . Note that heating is greater for larger particles undergoing vertical atmospheric entry. The high density of metal micrometeoroids leads to higher peak temperatures than silicate-dominated particles. **b,** Altitude at which peak temperature is reached for I-type spherules (numbers on curves, in km) with an entry

velocity of  $12 \text{ km s}^{-1}$ . Particles with higher entry velocity have similar peak altitudes since mass loss through evaporation leads to increased deceleration. **c,** The final radii of I-type particles (numbers on curves, in  $\mu\text{m}$ ) after deceleration at entry velocities of  $12 \text{ km s}^{-1}$  (black) and  $18 \text{ km s}^{-1}$  (grey). Greater mass loss occurs at higher entry velocities, entry angles and particle sizes. Mass loss occurs by surface evaporation of the exterior oxide melt.

**Extended Data Table 1 | Representative analyses of micrometeorite magnetite, wüstite and metal**

Sample	ID	1.1(2)	1.1(2)	1.1(2)	1.1(2)	1.1(2)	1.3(a)*
Fe (at.%)		43.46	43.82	42.74	50.07	50.35	90.16
Ni (at.%)		n.a.	n.a.	n.a.	n.a.	n.a.	6.41
O (at.%)		56.54	56.18	57.26	49.93	49.65	3.43
Formula		Fe <sub>3</sub> O <sub>4</sub>	Fe <sub>3</sub> O <sub>4</sub>	Fe <sub>3</sub> O <sub>4</sub>	FeO	FeO	FeNi + FeO
Mineral		magnetite	magnetite	magnetite	wüstite	wüstite	metal + wüstite <sup>†</sup>

n.a., not analysed.

\*This analysis comes from the metal particle rimmed by wüstite in Fig. 1e.

<sup>†</sup>Owing to the small size of the metal particles, analyses of these invariably included some surrounding wüstite in the total.

# A rapid burst in hotspot motion through the interaction of tectonics and deep mantle flow

Rakib Hassan<sup>1</sup>, R. Dietmar Müller<sup>1</sup>, Michael Gurnis<sup>2</sup>, Simon E. Williams<sup>1</sup> & Nicolas Flament<sup>1</sup>

**Volcanic hotspot tracks featuring linear progressions in the age of volcanism are typical surface expressions of plate tectonic movement on top of narrow plumes of hot material within Earth's mantle<sup>1</sup>. Seismic imaging reveals that these plumes can be of deep origin<sup>2</sup>—probably rooted on thermochemical structures in the lower mantle<sup>3–6</sup>. Although palaeomagnetic and radiometric age data suggest that mantle flow can advect plume conduits laterally<sup>7,8</sup>, the flow dynamics underlying the formation of the sharp bend occurring only in the Hawaiian–Emperor hotspot track in the Pacific Ocean remains enigmatic. Here we present palaeogeographically constrained numerical models of thermochemical convection and demonstrate that flow in the deep lower mantle under the north Pacific was anomalously vigorous between 100 million years ago and 50 million years ago as a consequence of long-lasting subduction systems, unlike those in the south Pacific. These models show a sharp bend in the Hawaiian–Emperor hotspot track arising from the interplay of plume tilt and the lateral advection of plume sources. The different trajectories of the Hawaiian and Louisville hotspot tracks arise from asymmetric deformation of thermochemical structures under the Pacific between 100 million years ago and 50 million years ago. This asymmetric deformation waned just before the Hawaiian–Emperor bend developed, owing to flow in the deepest lower mantle associated with slab descent in the north and south Pacific.**

The sharp bend in the Hawaiian–Emperor hotspot track in the north Pacific has been interpreted as the result of a sudden change in Pacific plate motion over a fixed plume. In contrast, the Louisville hotspot track in the south Pacific features a more gradual arc through the period of the bend<sup>9</sup>. Palaeomagnetic and radiometric age data suggest that the Hawaiian plume underwent a phase of rapid 11°–15° southward motion between 81 million years (Myr) ago and 47 Myr ago<sup>7</sup>, while the Louisville plume remained within 5°–7° of present-day latitude between 70 Myr ago and 50 Myr ago<sup>8</sup>, suggesting that the plumes are moving independently<sup>8</sup>. Seismic imaging reveals that these are robust plumes deeply rooted in the mantle<sup>2</sup> and deep plumes are likely to originate from thermochemical structures<sup>3–6</sup>. Moreover, the spatial correlation of hotspot locations with the edges of large low-shear-velocity provinces (LLSVPs)<sup>6</sup> suggests that these plumes are likely to be rooted on thermochemical ridges<sup>3,10</sup>. Numerical models of thermochemical structures suggest that their edges should be mobile<sup>11,12</sup>, in contrast to the longer-term stability implied by the correlation of the edges of LLSVPs and reconstructed volcanic features<sup>13,14</sup>.

A combination of plume motions derived from backward advection of mantle flow with a model of plate motions that explicitly incorporated a major change in relative plate motions occurring 52–43 Myr ago has been proposed to reproduce the observed Hawaiian–Emperor track<sup>15</sup>. These earlier flow models result in a substantial misfit of the predicted trail before 65 Myr ago<sup>15</sup>, and predict a gradual slowdown of the Hawaiian plume motion, compared to the abrupt slowdown occurring around the time of formation of the Hawaiian–Emperor

bend, as indicated by palaeomagnetic data. Moreover, Pacific–Farallon relative plate motion changes contemporaneous with the Hawaiian–Emperor bend were insignificant and gradual<sup>7,16</sup>, suggesting that the bend is primarily a consequence of rapid southward migration of the Hawaiian plume<sup>7,17,18</sup>.

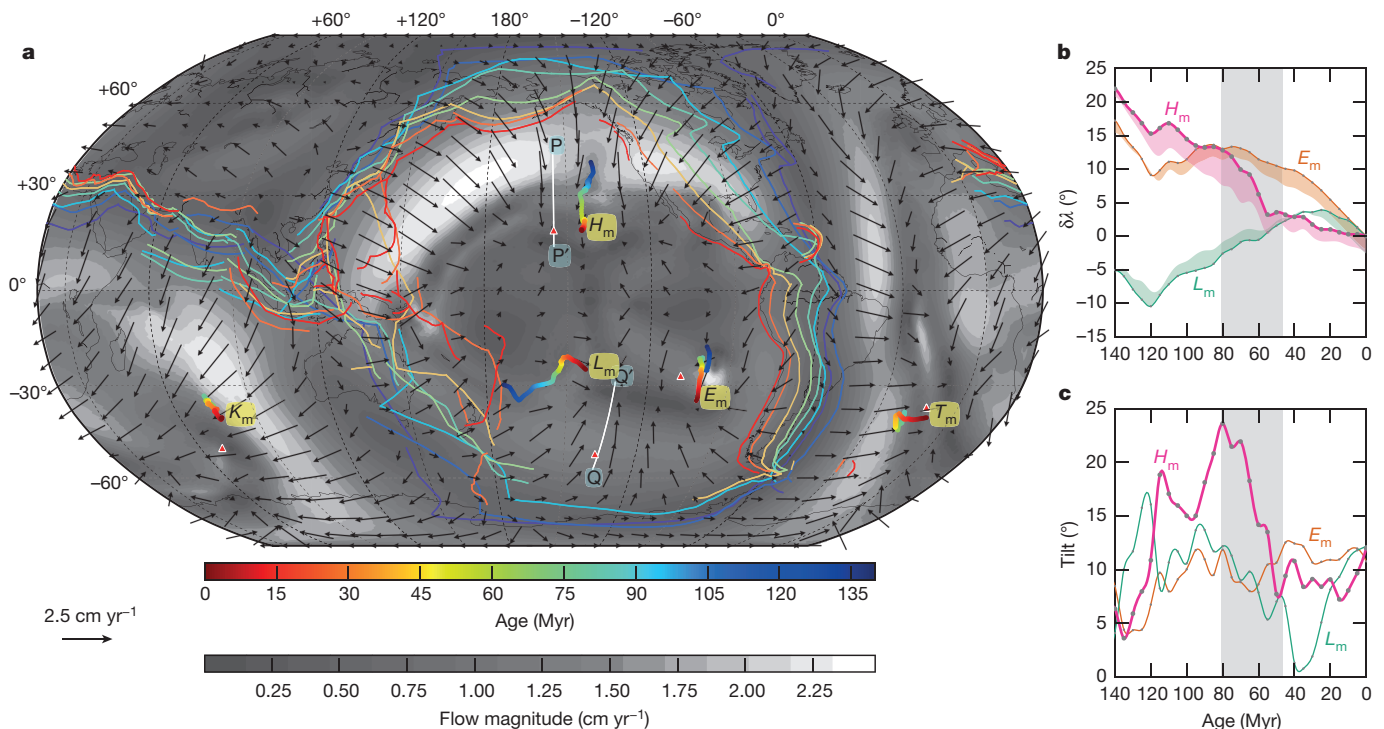
A capture–release mechanism has also been proposed<sup>17</sup>, in which the Hawaiian plume may have been captured by the fast-spreading Kula–Pacific ridge system, situated to the north of the Detroit seamount during its construction. In this scenario, vigorous shallow mantle flow induced a strong plume tilt starting at mid-mantle depth, before cessation of spreading between 56 Myr ago and 47 Myr ago resulted in a change in mantle flow regime. Consequently, the plume conduit rapidly returned to its original position, resulting in the Hawaiian–Emperor bend<sup>17</sup>. However, an assumption implicit in this scenario is that the base of the plume remains fixed, relative to lower mantle structures, during the change in flow regime.

Instead, here we advance the hypothesis that between about 100 Myr ago and 50 Myr ago deep mantle flow beneath the north Pacific was anomalously vigorous, a consequence of robust and long-lasting subduction systems, unlike those in the south Pacific. This asymmetric flow in the deepest lower mantle under the north and south Pacific explains the rapid southward motion of the Hawaiian plume and the contrasting sluggish latitudinal motion of the Louisville plume, resulting in sharp differences in the geometry of hotspot tracks.

We follow the trajectory and tilt of plumes over the past 230 Myr with palaeogeographically constrained, spherical numerical models of convection<sup>10,11,19</sup>. We impose kinematic boundary conditions, consistent with plate motions, and assimilate thermal models of shallow slabs at convergent plate margins (Methods). Descending slab material deforms a basal layer of anomalous density<sup>20,21</sup> (Methods), covering the core–mantle boundary (CMB), where the slab descent rates observed in our models<sup>10</sup> are consistent with estimates inferred from global mantle tomography<sup>22</sup>. Viscosity in the deepest lower mantle, below the viscosity peak<sup>23,24</sup>, is low, owing to elevated temperatures in the thermal boundary layer above the CMB<sup>21,25</sup>, and thus flow velocities there are expected to be higher than those in the ambient lower mantle. This is a consistent feature of our models, in agreement with observations from earlier flow models<sup>26</sup>. Consequently, within the first 50 Myr of model time, the dense material above the CMB deforms, ultimately resembling the seismologically observed LLSVPs in the lower mantle (Methods, Extended Data Fig. 1). We refer to these dense thermochemical structures above the CMB in our models as LLSVPs. Fully dynamic plumes nucleate primarily around the LLSVPs and plume eruption locations are highly correlated to the reconstructed eruption locations of large igneous provinces<sup>10</sup>. The models are characterized roughly by as many plumes at the present day as can be inferred to be of deep origin from tomographic studies<sup>2,27</sup>.

Subduction history has been suggested to have a first-order role in shaping the geometry of LLSVPs<sup>10–12,28</sup>. Here we compute poloidal flow, associated with the buoyancy resulting from subduction, in the

<sup>1</sup>EarthByte Group, School of Geosciences, University of Sydney, Sydney, New South Wales 2006, Australia. <sup>2</sup>Seismological Laboratory, California Institute of Technology, Pasadena, California 91125, USA.



**Figure 1 | Time evolution of model plume trajectories in a model of thermochemical convection.** See Extended Data Table 2, case M3. **a**, The magnitude of the time-averaged mean poloidal flow (see text) is shown by grey shading and the corresponding flow directions are shown by black arrows. Subduction zones over the past 140 Myr are plotted at 20-Myr intervals, coloured by age. Red triangles indicate the present-day location of observed plumes of deep origin at the surface. The corresponding model plume trajectories for Hawaii ( $H_m$ ), Louisville ( $L_m$ ), Easter ( $E_m$ ), Kerguelen ( $K_m$ ) and Tristan ( $T_m$ ), at a depth of 350 km, are coloured by age. These trajectories generally reflect the mean poloidal flow. Model plumes

deep lower mantle to evaluate the temporal evolution of the edges of LLSVPs. We compute mean poloidal flow within a 300-km-thick shell above the CMB and derive a time average of this evolving mean poloidal flow over the past 140 Myr of modelled time (Methods), highlighting regions that undergo strong coherent poloidal flow relative to regions where it is comparatively weak and diffuse. Over the past 140 Myr, subduction zones bounding the north Pacific migrate oceanwards (Fig. 1a), so the deep mantle here is dominated by a strong coherent flow. In contrast, flow in the south Pacific is diffuse and the darker region under the central Pacific marks the mean areal extent of the Pacific LLSVP, where mean poloidal flow is weakest (Fig. 1a).

The southward trajectory of the Hawaiian plume is nearly perpendicular to the strike of the northern edge of the Pacific LLSVP, marked by zones of quasi-stagnation (Fig. 1a). Southeastward flow beneath the northwest Pacific converges with southwestward flow from the northeast Pacific, resulting in a strong net southward flow that reaches a maximum along a corridor coincident with the location of the present-day Hawaiian plume. The edge of the Pacific LLSVP recedes southward along this corridor by more than 15° over the past 100 Myr (Extended Data Fig. 2 and Supplementary Video 1), making it a unique geodynamical setting. The pulse of southward poloidal flow in the north Pacific during the period 100–50 Myr ago is counterbalanced by the northeasterly poloidal flow under the southwest Pacific that strengthens after about 85 Myr ago (Extended Data Fig. 2 and Supplementary Video 1)—resulting in the transition from a phase of rapid southward motion of the model Hawaiian plume,  $H_m$ , to a much slower regime at about 50 Myr ago (Fig. 1b, Extended Data Fig. 3 and Supplementary Video 2).

A southward velocity of around 3 cm yr<sup>-1</sup> in the deep lower mantle at 90 Myr ago (Fig. 2a, bottom row) translates to over 5 cm yr<sup>-1</sup>

in the Indo-Atlantic are shorter-lived and less mobile than those in the Pacific. See Fig. 2 for time-dependent profiles along PP' and QQ'. **b**, The latitudinal deviation,  $\delta\lambda$ , of model plumes from their present-day locations at a depth of 350 km. The coloured shading shows the latitudinal offset between the locations of plume conduits at depths of 350 km and 1,500 km. Between 81 Myr ago and 47 Myr ago (shaded grey rectangle),  $H_m$  rapidly moves southward by about 10°. **c**, Model plume tilt, as outlined in the Methods. Between about 81 Myr ago and 47 Myr ago, tilt accumulated by  $H_m$  over an earlier period of coherent southward motion decreases sharply.

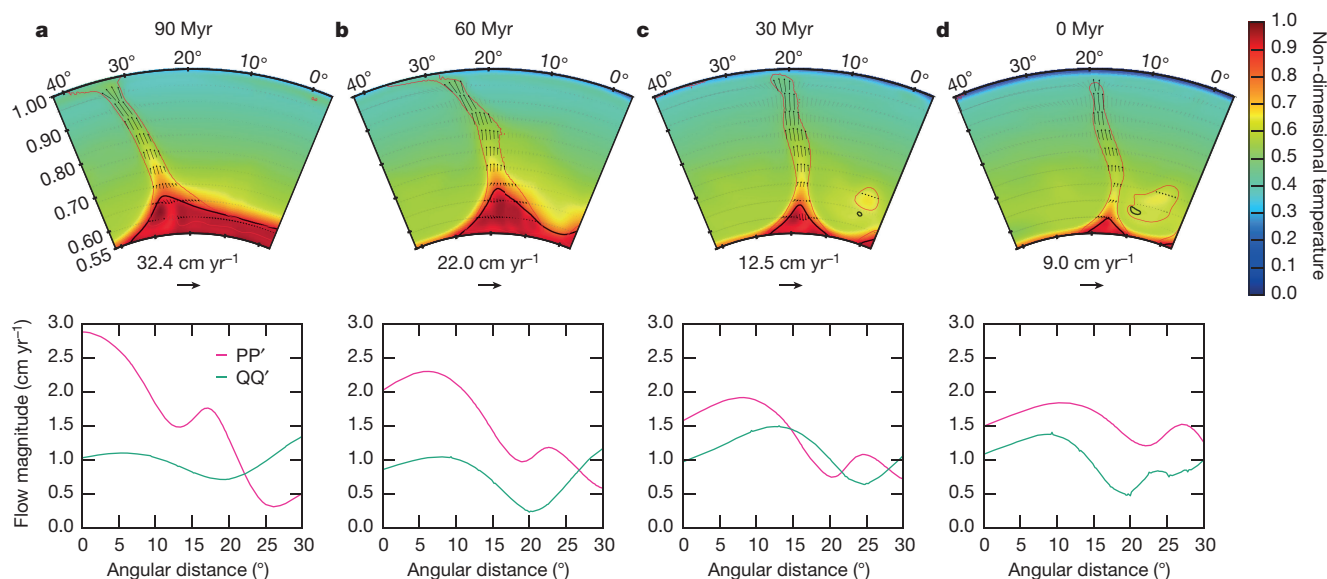
near the surface, consistent with palaeomagnetic estimates<sup>7</sup>. Over the same period, the edge of the LLSVP in the south Pacific—marked by the trough in velocity (Fig. 2a–d, bottom row)—remains comparatively stable. The Louisville plume is thus isolated from the kind of rapid burst and slowdown of motion that caused the sharp Hawaiian–Emperor bend.

A dynamical explanation for the sharp bend in the Hawaiian–Emperor track requires an abrupt slowdown of the plume at the time of the bend's formation<sup>15</sup> that cannot be accounted for by deep lower-mantle flow alone. Ascent speed within plume conduits is low in the vicinity of the viscosity peak in the lower mantle and is comparable to rates of horizontal advection, whereas in the shallow lower mantle and above, the rate of ascent dominates the motion of a plume conduit.

Consequently, the lateral advection of plume sources rooted on chemical ridges in the deep lower mantle is the primary factor controlling plume tilt. Between about 81 Myr ago and 47 Myr ago, the surface location of  $H_m$  moves southward by about 10°, followed by a sharp slowdown, whereas during this period, the base of the plume moves only about 6° (Fig. 1b). The remaining ~4° of motion near the surface arises from the slowdown of southward motion of the base of  $H_m$ . Consequently, the latitudinally tilted plume rapidly straightens, resulting in an abrupt slowdown in hotspot motion near the surface at about 47 Myr ago (Fig. 1c, Extended Data Fig. 3 and Supplementary Video 2). Predicted hotspot tracks are nearly identical to the observed Hawaiian–Emperor track (Fig. 3a) and the latitudinal motion of  $H_m$  corresponds to the much faster southward motion of the Hawaiian plume before the Hawaiian–Emperor bend occurs (Fig. 3b), consistent with palaeomagnetic data<sup>7</sup>.

Numerical experiments computed over a range of uncertain parameters (Methods, Extended Data Tables 1 and 2) demonstrate a





**Figure 2 | Longitudinal cross-sections.** Cross-sections through  $H_m$  (see Fig. 1a) show its southward motion and the evolution of tilt that arises from the offset of the base of the plume relative to conduit location near the surface (see Fig. 1b and c). Ascent speeds within model plume conduits wane towards the present-day values as plumes mature. This is consistent with large volumes of lava being emplaced on Earth's surface when a plume-head erupts, followed by weaker magmatic activity associated with a waning plume-tail. **a**, Temperature profile along a constant meridian, through  $H_m$ , at age 90 Myr, for latitudinal extent shown.

The temperature field away from radial layer averages in the panel is contoured in red by the depth-averaged standard deviation of radial layer temperature values, thus outlining the plume conduit. The thick black contour marks 75% anomalous chemical concentration. Velocity vectors outside the plume conduit are shown in lighter shading. The bottom panel shows the magnitude of mean poloidal flow in a 300-km-thick shell above the CMB, at age 90 Myr, along profiles PP' and QQ' (see Fig. 1a). **b–d**, As for **a**, but for ages as given.

consistent pattern of strong time-averaged poloidal flow in the north Pacific compared to that in the south (Extended Data Fig. 4). Rapid southward migration of  $H_m$  is observed when the density contrast of

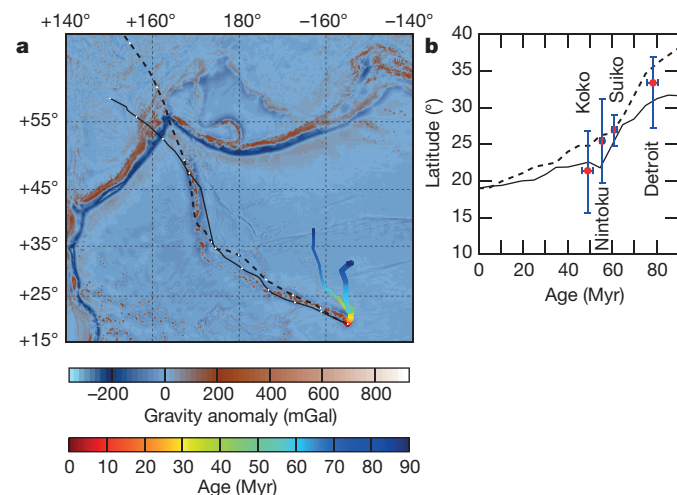
the LLSVP material ranges between 2% and 2.5%; for larger density contrasts (Extended Data Fig. 4f) and when the LLSVP material is intrinsically more viscous (Extended Data Fig. 4e), the areal extent of the Pacific LLSVP is greater and the plume motion is smaller. Plume mobility could thus be a criterion to constrain the density contrast of LLSVP material.

We demonstrate that for a range of reasonable parameter values, the Pacific LLSVP deforms asymmetrically between 100 Myr ago and 50 Myr ago and that predicted hotspot tracks explain the formation of the sharp Hawaiian–Emperor bend, without requiring a major change in plate motion. We note that our present models do not capture the potential influences of a waning ridge system in the vicinity of a plume and that more sophisticated convection models—featuring self-consistent ridges and transform boundaries—are necessary to test the dynamic plausibility of a capture–release mechanism<sup>17</sup> for plumes rooted on mobile lower mantle structures. Our results imply that the Hawaiian plume may have been active since approximately 140 Myr ago, longer than previously recognized, and that the remnants of this activity would have been subducted beneath the Kamchatka peninsula<sup>29</sup>. Even though there is robust evidence for the overall stability of the African LLSVP for times when Africa was relatively stationary and distant from migrating subduction zones<sup>30</sup>, our results demonstrate that a tectonic setting of fast-moving oceanic plates surrounded by a migrating, dynamic set of subduction zones can lead to substantial LLSVP deformation and hotspot mobility.

**Online Content** Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

**Received 3 November 2015; accepted 9 February 2016.**

1. Morgan, W. J. Convection plumes in the lower mantle. *Nature* **230**, 42–43 (1971).
2. French, S. W. & Romanowicz, B. Broad plumes rooted at the base of the Earth's mantle beneath major hotspots. *Nature* **525**, 95–99 (2015).
3. Jellinek, A. M. & Manga, M. The influence of a chemical boundary layer on the fixity, spacing and lifetime of mantle plumes. *Nature* **418**, 760–763 (2002).



**Figure 3 | Predicted hotspot tracks.** **a**, The thick multicoloured trajectory represents the motion of the Hawaiian plume, based on the motion of  $H_m$  (see Fig. 1). The thin multicoloured trajectory represents the same from a similar model (see Extended Data Table 2, case M4), where the plate reconstruction is based on a reference frame constrained by lower-mantle slab remnants (Methods). The solid and dashed black lines represent predicted surface hotspot tracks, computed from the thick and thin multicoloured trajectories of  $H_m$ , respectively (Methods). White diamonds along these hotspot tracks mark 10-Myr intervals. The background shows gravity anomaly to reveal the Hawaiian–Emperor seamount chain. **b**, The solid and dashed black lines show palaeolatitudes corresponding to the thick and thin multicoloured trajectories in **a**, respectively. Red circles are inferred palaeolatitudes for the Emperor seamounts and vertical error bars represent 95% confidence limits<sup>7</sup>. See Tarduno *et al.*<sup>7</sup> for a description of horizontal error bars and lists of data sources.

4. Lin, S.-C. & van Keken, P. E. Dynamics of thermochemical plumes: 2. Complexity of plume structures and its implications for mapping mantle plumes. *Geochem. Geophys. Geosyst.* **7**, Q03003 (2006).
5. Farnetani, C. Excess temperature of mantle plumes: the role of chemical stratification across D''. *Geophys. Res. Lett.* **24**, 1583–1586 (1997).
6. Thorne, M. S., Garnero, E. J. & Grand, S. P. Geographic correlation between hot spots and deep mantle lateral shear-wave velocity gradients. *Phys. Earth Planet. Inter.* **146**, 47–63 (2004).
7. Tarduno, J. A. *et al.* The Emperor seamounts: southward motion of the Hawaiian hotspot plume in Earth's mantle. *Science* **301**, 1064–1069 (2003).
8. Koppers, A. A. *et al.* Limited latitudinal mantle plume motion for the Louisville hotspot. *Nature Geosci.* **5**, 911–917 (2012).
9. Koppers, A. A., Duncan, R. A. & Steinberger, B. Implications of a nonlinear <sup>40</sup>Ar/<sup>39</sup>Ar age progression along the Louisville seamount trail for models of fixed and moving hot spots. *Geochem. Geophys. Geosyst.* **5**, Q06L02 (2004).
10. Hassan, R., Flament, N., Gurnis, M., Bower, D. J. & Müller, D. Provenance of plumes in global convection models. *Geochem. Geophys. Geosyst.* **16**, 1465–1489 (2015).
11. Bower, D. J., Gurnis, M. & Seton, M. Lower mantle structure from paleogeographically constrained dynamic Earth models. *Geochem. Geophys. Geosyst.* **14**, 44–63 (2013).
12. McNamara, A. K. & Zhong, S. Thermochemical structures beneath Africa and the Pacific Ocean. *Nature* **437**, 1136–1139 (2005).
13. Burke, K., Steinberger, B., Torsvik, T. H. & Smethurst, M. A. Plume generation zones at the margins of large low shear velocity provinces on the core–mantle boundary. *Earth Planet. Sci. Lett.* **265**, 49–60 (2008).
14. Torsvik, T. H. *et al.* Deep mantle structure as a reference frame for movements in and on the Earth. *Proc. Natl Acad. Sci. USA* **111**, 8735–8740 (2014).
15. Steinberger, B., Sutherland, R. & O'Connell, R. J. Prediction of Emperor–Hawaii seamount locations from a revised model of global plate motion and mantle flow. *Nature* **430**, 167–173 (2004).
16. Wright, N. M., Müller, R. D., Seton, M. & Williams, S. E. Revision of paleogene plate motions in the Pacific and implications for the Hawaiian–Emperor bend. *Geology* **43**, 455–458 (2015).
17. Tarduno, J., Bunge, H.-P., Sleep, N. & Hansen, U. The bent Hawaiian–Emperor hotspot track: inheriting the mantle wind. *Science* **324**, 50–53 (2009).
18. Tarduno, J. A. On the motion of Hawaii and other mantle plumes. *Chem. Geol.* **241**, 234–247 (2007).
19. Bower, D. J., Gurnis, M. & Flament, N. Assimilating lithosphere and slab history in 4-D Earth models. *Phys. Earth Planet. Inter.* **238**, 8–22 (2015).
20. Garnero, E. J. & McNamara, A. K. Structure and dynamics of Earth's lower mantle. *Science* **320**, 626–628 (2008).
21. Lay, T., Williams, Q. & Garnero, E. J. The core–mantle boundary layer and deep Earth dynamics. *Nature* **392**, 461–468 (1998).
22. van der Meer, D. G., Spakman, W., van Hinsbergen, D. J., Amaru, M. L. & Torsvik, T. H. Towards absolute plate motions constrained by lower-mantle slab remnants. *Nature Geosci.* **3**, 36–40 (2010).
23. Steinberger, B. & Calderwood, A. R. Models of large-scale viscous flow in the Earth's mantle with constraints from mineral physics and surface observations. *Geophys. J. Int.* **167**, 1461–1481 (2006).
24. van Keken, P. E., Yuen, D. A. & van den Berg, A. P. Implications for mantle dynamics from the high melting temperature of perovskite. *Science* **264**, 1437–1439 (1994).
25. Olson, P., Schubert, G. & Anderson, C. Plume formation in the D''-layer and the roughness of the core–mantle boundary. *Nature* **327**, 409–413 (1987).
26. Steinberger, B. Plumes in a convecting mantle: models and observations for individual hotspots. *J. Geophys. Res. Solid Earth* **105**, 11127–11152 (2000).
27. Montelli, R. *et al.* Finite-frequency tomography reveals a variety of plumes in the mantle. *Science* **303**, 338–343 (2004).
28. Zhong, S. & Rudolph, M. L. On the temporal evolution of long-wavelength mantle structure of the Earth since the early Paleozoic. *Geochem. Geophys. Geosyst.* **16**, 1599–1615 (2015).
29. Portnyagin, M., Savelyev, D., Hoernle, K., Hauff, F. & Garbe-Schönberg, D. Mid-Cretaceous Hawaiian tholeiites preserved in Kamchatka. *Geology* **36**, 903–906 (2008).
30. Torsvik, T. H., Steinberger, B., Cocks, L. R. M. & Burke, K. Longitude: linking Earth's ancient surface to its deep interior. *Earth Planet. Sci. Lett.* **276**, 273–282 (2008).

**Supplementary Information** is available in the online version of the paper.

**Acknowledgements** M.G. was supported by the NSF (awards EAR-1161046 and EAR-1247022). R.D.M. and N.F. were supported by an ARC grant (IH130200012). This research was undertaken with the assistance of resources from the National Computational Infrastructure (NCI), which is supported by the Australian Government.

**Author Contributions** R.H. and R.D.M. developed the concept of the study. R.H. and M.G. designed the numerical experiments and developed the technical aspects of the study. All authors contributed both intellectually and to the writing of the paper.

**Author Information** Reprints and permissions information is available at [www.nature.com/reprints](http://www.nature.com/reprints). The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to R.H. ([rakib.hassan@sydney.edu.au](mailto:rakib.hassan@sydney.edu.au)).

## METHODS

**The numerical model.** We consider models of thermochemical convection within Earth's mantle under the extended-Boussinesq approximation<sup>31</sup> in a spherical shell, with depth-dependent thermodynamic properties and temperature- and depth-dependent rheology. The equations for the conservation of mass, momentum and energy are solved using Citcoms<sup>32</sup> as a finite-element problem, which has been modified to allow for progressive assimilation of surface plate motion and inferred slab material based on global plate reconstructions<sup>19</sup>. Our choice of model parameters and underlying assumptions have been outlined in earlier work<sup>10,11</sup>. Extended Data Table 1 lists parameters held fixed across all model cases and additional details can be found in our earlier work<sup>10</sup>.

**Model setup.** The spherical shell representing Earth's mantle is decomposed into about 12.6 million mesh elements. Radial mesh refinement provides a vertical resolution of about 15 km and about 27 km near the top and bottom boundary layers, respectively. We assume an a priori mantle adiabat augmented by thermal boundary layers. The top and the bottom thermal boundary layers each encompass a temperature drop of 1,225 K and the initial adiabatic temperature profile has a potential temperature of 1,525 K. A non-dimensional internal heat generation rate of 100 is applied in all model cases. We compute a reference profile for thermal expansion,  $\alpha$ , based on analytical parameterizations given in ref. 33, using the a priori mantle temperature profile.

We use piecewise Arrhenius laws to describe the variation of viscosity with temperature, depth and composition in Earth's mantle, which takes the following non-dimensional form:

$$\eta(T, r) = A(r)\eta_c \exp\left(\frac{E_a(r) + (1-r)V_a(r)}{T + T_{\text{off}}} - \frac{E_a(r) + (1-r_{\text{inner}})V_a(r)}{1 + T_{\text{off}}}\right) \quad (1)$$

where  $\eta$  is the viscosity,  $T$  is the temperature,  $r$  is the radius,  $A$  is the pre-exponential parameter,  $\eta_c$  is the intrinsic composition-dependent pre-factor,  $E_a$  is the activation energy,  $V_a$  is the activation volume,  $T_{\text{off}}$  is the temperature offset and  $r_{\text{inner}}$  is the radius at the CMB. The second term within the exponential ensures that it reduces to 1 when  $T = 1$  and  $r = r_{\text{inner}}$ . For the lower mantle, we use a dimensional activation energy of 320 kJ mol<sup>-1</sup> and activation volume of  $6.7 \times 10^{-6}$  m<sup>3</sup> mol<sup>-1</sup>, corresponding to non-dimensional units of 11 and 26, respectively, comparable to estimates in ref. 34. However, such viscosity parameterizations lead to large viscosity variations, causing numerical difficulties. To limit the viscosity contrast to three orders of magnitude, we adjust the pre-exponential parameter  $A(r)$  and the temperature offset  $T_{\text{off}}$  (ref. 35). The resulting viscosity profile is similar to the preferred viscosity profiles of ref. 23. Additional details on model setup can be found in our earlier work<sup>10</sup>.

**Initial and boundary conditions.** We apply kinematic surface boundary conditions that are a function of both the relative plate motion model and the absolute plate motion model. We use the relative plate motion from ref. 36, updated for circum-Arctic regions as described in ref. 37. In most model cases we use the absolute plate motion model of ref. 30 for the period 0–70 Myr ago, and the absolute plate motion model of ref. 38 for the period 105–200 Myr ago, with an interpolation between 70 Myr ago and 105 Myr ago. We refer to this hybrid absolute plate motion model as APM1 in Extended Data Table 2. We also test a kinematic scenario using the same relative plate motion model combined with an absolute plate motion model constrained by lower mantle slab remnants<sup>22</sup>, which we refer to as APM2 in Extended Data Table 2. Surface velocities derived from these global plate tectonic reconstructions are obtained in 1-Myr intervals with a linear interpolation in between.

We constructed a thermal model of slabs from the reconstructed location and age of oceanic lithosphere at convergent plate margins. These slabs are initially inserted from the surface to a depth of 1,200 km. Slabs that appear during modelled geological time are progressively inserted into the upper mantle on the basis of their age of appearance. This imposed thermal structure is blended with the dynamically evolving temperature field at each timestep—see Bower *et al.*<sup>11,19</sup> for a more detailed description of the progressive data assimilation method. This simple approach captures the essential aspects of subduction by injecting slabs with realistic thickness and mass flux, without requiring complex rheological laws to model the physics of plate boundaries in our models.

To evaluate the influence of parameter values and initial conditions we have computed a range of models in which we varied the density contrast of the dense chemical layer above the CMB, the intrinsic viscosity contrast of the dense material, the initial geometry of the dense layer and the plate reconstruction that dictates surface plate velocities and the location of subduction zones. Extended Data Table 2 lists the model cases presented in this study. In most model cases, an initially uniform layer of anomalously dense material about 100 km thick covers the CMB. However, before the Mesozoic era, the dense LLSVP material may have already been displaced and deformed by slabs<sup>11,39</sup>. To test the influence of

an already thickened dense layer on model outcomes, we present a model case (M5 in Extended Data Table 2) where the dense material is confined to two slightly elongated domes—resembling present-day LLSVPs—in the initial condition. These domes are about 1,000 K hotter than the ambient mantle<sup>11</sup> and have a relief of about 450 km, with their combined volume comparable to estimates of the volume of LLSVPs<sup>13</sup>.

**Poloidal flow in the lower mantle.** The velocity field  $\mathbf{u}$  in our global models is divergence free:

$$\nabla \cdot \mathbf{u} = 0 \quad (2)$$

This allows us to decompose the tangential velocity field on each radial shell of the spherical mesh into poloidal and toroidal components. The poloidal flow originates from buoyancy-driven convergent or divergent flows, such as at spreading centres and subduction zones. Following from Helmholtz's theorem on spheres<sup>40</sup>:

$$\mathbf{V}_s = \nabla_s \Phi + \nabla_s \times (\Psi \mathbf{r}) \quad (3)$$

where  $\mathbf{V}_s$  is the tangential velocity on a radial shell of the computational domain,  $\nabla_s$  is the gradient operator on a spherical shell,  $\Phi$  is the poloidal potential,  $\Psi$  is the toroidal potential and  $\mathbf{r}$  is the radial unit vector. See ref. 40 for more details. The poloidal potential  $\Phi$  is obtained as follows:

$$\Phi = \nabla_s^{-2} (\nabla_s \cdot \mathbf{V}_s) \quad (4)$$

Subsequently, the poloidal flow field  $\mathbf{V}_p$  on a radial shell is obtained as:

$$\mathbf{V}_p = \nabla_s \Phi \quad (5)$$

**Model hotspot tracks and plume tilt.** A detailed account of a scheme for detecting plumes in global mantle convection models is given in ref. 10. We process model outputs at 5-Myr intervals, since the transit time of a plume head from nucleation to eruption under the base of the lithosphere is always  $> 5$  Myr. We compute a set of extant plume conduit locations,  $\mathbf{S}_t = \{P_1, \dots, P_n\}$ , at a depth of 350 km, based on the plume detection scheme, where subscript  $n$  denotes the total number of plumes detected at a given model time  $t$ . Age-progressive model hotspot tracks in the mantle frame of reference are subsequently identified by binning conduit locations from the ensemble set  $\mathbf{S} = \{\mathbf{S}_{t_0}, \dots, \mathbf{S}_{t_m}\}$  based on spatiotemporal proximity, where  $t_0$  is the model time when the first model plumes appear and  $t_m$  is the model time corresponding to the present day.

A fourth-order Runge–Kutta scheme is used to compute streamlines passing through each  $P_i \in \mathbf{S}_t$  and spanning mantle depths of 350–1,500 km that mark the topological skeletons of extant plume conduits at a given model time. We limit the depth extent of these streamlines to 1,500 km, since plumes are rooted on chemical ridges that can reach more than about 1,000 km above the CMB. Plume tilt is then computed based on  $P_i$  and its corresponding location at 1,500 km depth, relative to the radial normal.

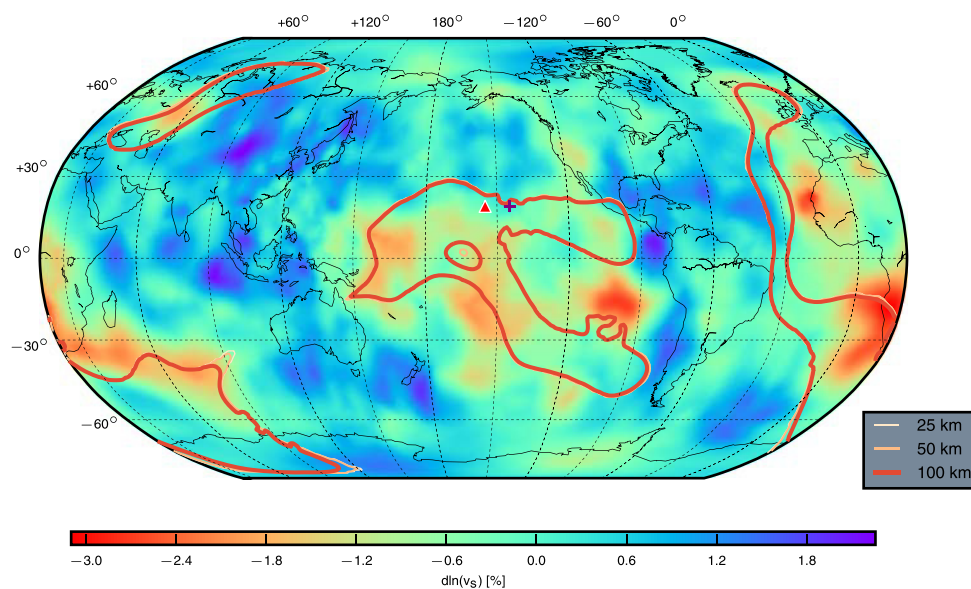
**Predicted hotspot tracks.** The motion of a model plume in the mantle frame of reference can be represented implicitly by a sequence of finite rotations on the sphere, decoupling it from explicit spatial locations. By applying such motion trajectories obtained from convection models—such as  $H_m$  (Fig. 1a)—to the corresponding observed plume (Hawaii), a direct comparison can be made between the predicted and the observed hotspot tracks.

**Comparison of model LLSVPs with tomography.** Extended Data Fig. 1 shows a comparison of present-day model LLSVPs for case M3 (Extended Data Table 2) with the *Savani* tomography model<sup>41</sup>. The first-order shapes of both the African and Pacific LLSVPs are largely in agreement with our model. The western extension of the Pacific LLSVP is much smaller in north–south extent than in the centre, consistent with our model. The Pacific LLSVP in the east is broken up into blobs, where the details are a little different compared to our model, but it shows that in the east this feature is more discontinuous. However, model LLSVP edges in the vicinity of the actual and model Hawaiian plumes agree well with the tomography (Extended Data Fig. 1).

In the African hemisphere, the tomography model shows the long east–west-oriented eastern arm of our African model LLSVP in the southern central Indian Ocean, thickening to the west as it straddles South Africa, and a more discontinuous northern extension where we have a narrow arm stretching towards Iceland. The 'Perm Anomaly'<sup>42</sup> beneath present-day western Siberia appears as a rounded feature in the tomography model, as opposed to the more elongate corresponding feature in our model—though both are isolated and separated from the northern part of the African LLSVP. This comparison suggests that our models do a reasonable job of capturing lower mantle dynamics—especially considering the relative simplicity of the models—and it is not expected that model LLSVP edges would match higher-order geometric features found in mantle tomography models.

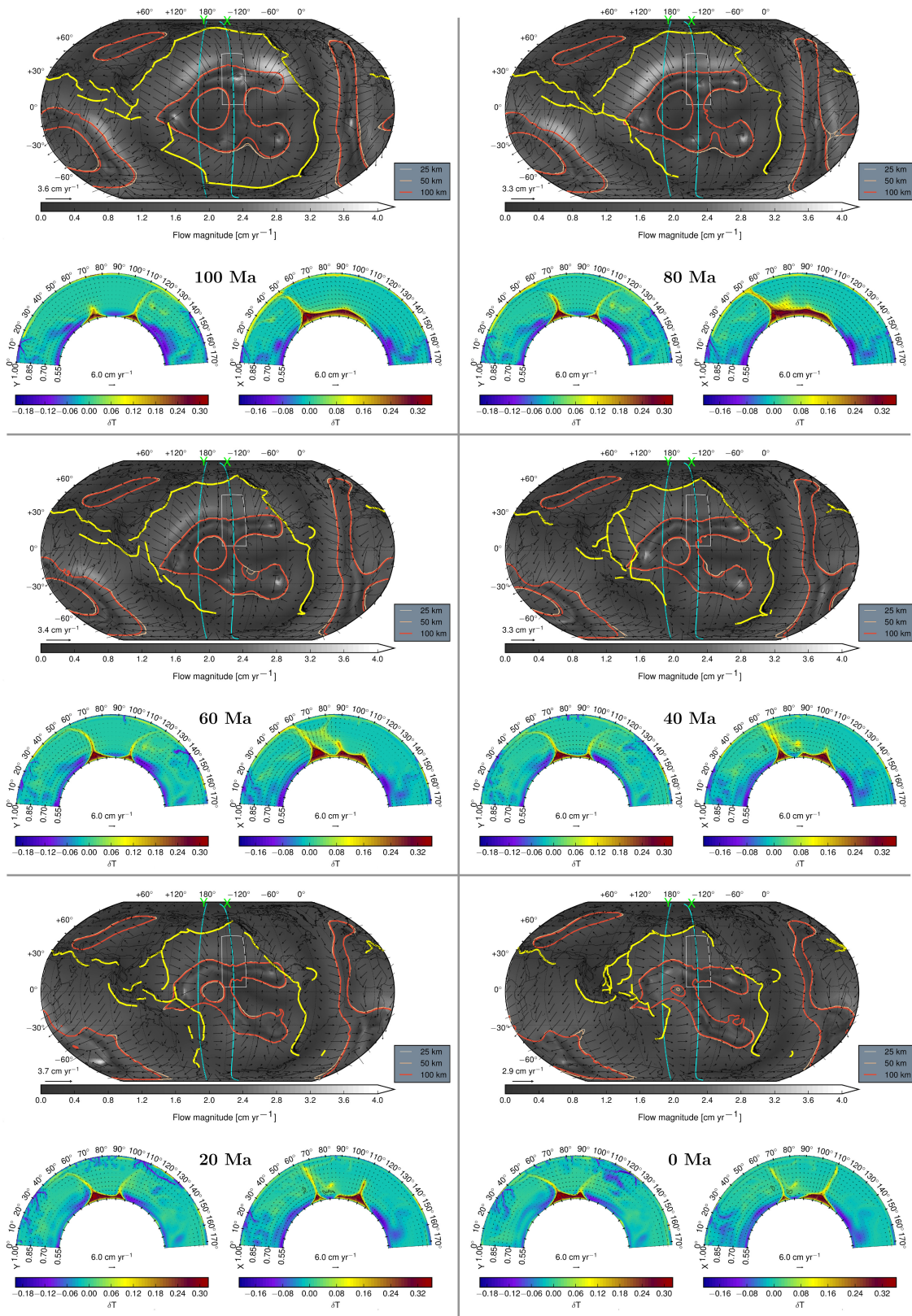
31. Christensen, U. R. & Yuen, D. A. Layered convection induced by phase transitions. *J. Geophys. Res.* **90**, 10291–10300 (1985).
32. Zhong, S., McNamara, A., Tan, E., Moresi, L. & Gurnis, M. A benchmark study on mantle convection in a 3-D spherical shell using CitcomS. *Geochem. Geophys. Geosyst.* **9**, Q10017 (2008).
33. Tosi, N., Yuen, D. A., de Koker, N. & Wentzcovitch, R. M. Mantle dynamics with pressure- and temperature-dependent thermal expansivity and conductivity. *Phys. Earth Planet. Inter.* **217**, 48–58 (2013).
34. Karato, S.-I. & Wu, P. Rheology of the upper mantle: a synthesis. *Science* **260**, 771–778 (1993).
35. Tackley, P. J. Effects of strongly variable viscosity on three-dimensional compressible convection in planetary mantles. *J. Geophys. Res.* **101**, 3311–3332 (1996).
36. Seton, M. *et al.* Global continental and ocean basin reconstructions since 200Ma. *Earth Sci. Rev.* **113**, 212–270 (2012).
37. Shephard, G. E., Müller, R. D. & Seton, M. The tectonic evolution of the Arctic since Pangea breakup: integrating constraints from surface geology and geophysics with mantle structure. *Earth Sci. Rev.* **124**, 148–183 (2013).
38. Steinberger, B. & Torsvik, T. H. Absolute plate motions and true polar wander in the absence of hotspot tracks. *Nature* **452**, 620–623 (2008).
39. McNamara, A. K. & Zhong, S. Thermochemical structures within a spherical mantle: superplumes or piles? *J. Geophys. Res.* **109**, B07402 (2004).
40. Backus, G. Poloidal and toroidal fields in geomagnetic field modeling. *Rev. Geophys.* **24**, 75–109 (1986).
41. Auer, L., Boschi, L., Becker, T., Nissen-Meyer, T. & Giardini, D. *Savani*: a variable resolution whole-mantle model of anisotropic shear velocity variations based on multiple data sets. *J. Geophys. Res.* **119**, 3006–3034 (2014).
42. Lekic, V., Cottaar, S., Dziewonski, A. & Romanowicz, B. Cluster analysis of global lower mantle tomography: a new class of structure and implications for chemical heterogeneity. *Earth Planet. Sci. Lett.* **357/358**, 68–77 (2012).





**Extended Data Figure 1 | Comparison of model LLSVPs with tomography.** The *Savani* tomography model<sup>41</sup>, showing shear velocity ( $v_s$ ) perturbations at 2,818 km depth. Contours of the 75% chemical concentration isosurface, at labelled heights above the CMB, show the

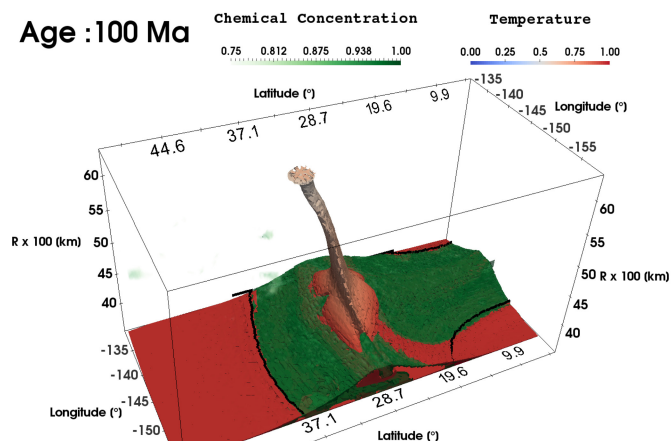
present-day shapes of the model LLSVPs in case M3 (Extended Data Table 2). The red triangle and purple cross symbols mark the locations of the actual and model Hawaiian plumes at present-day, respectively.



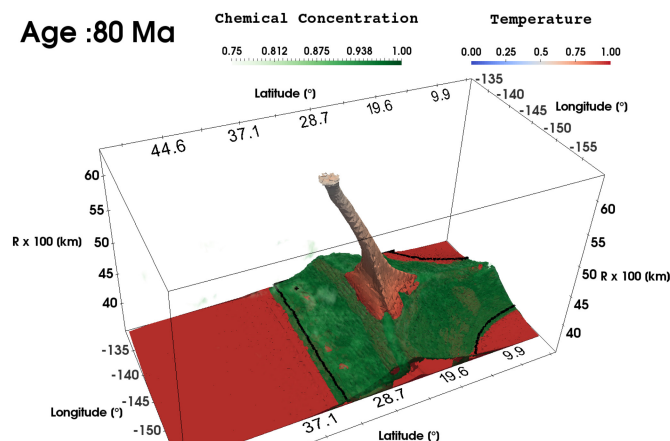
**Extended Data Figure 2 | Evolution of mean poloidal flow.** In each panel, the magnitude of mean poloidal velocity in a 300-km-thick shell above the CMB is shown in grey shading and corresponding flow directions are shown by black arrows at the age labelled, for case M3 (Extended Data Table 2). Edges of the model LLSVPs are marked by contours of the 75% chemical concentration isosurface, at labelled heights above the CMB. Subduction zones are shown in yellow and the white rectangular region

marks the extent of the three-dimensional plots in Extended Data Fig. 3. In the bottom row of each panel, cross-sections along cyan profiles through the Pacific LLSVP show the evolution of its edges driven by subduction-induced flow. Velocity vectors in these cross-sections have been clipped to 6 cm yr<sup>-1</sup> and the black contours show 75% chemical concentration.

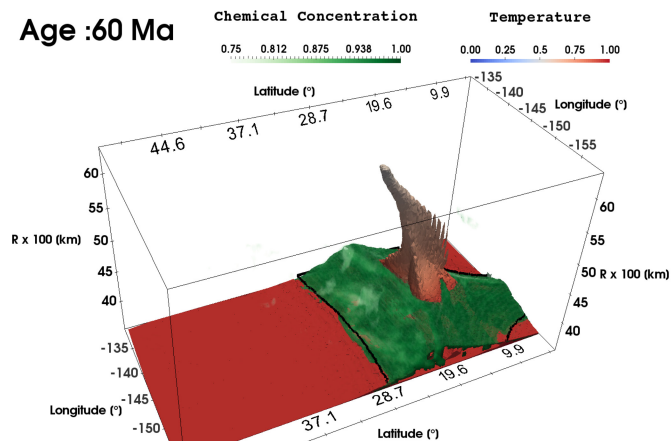
Age :100 Ma



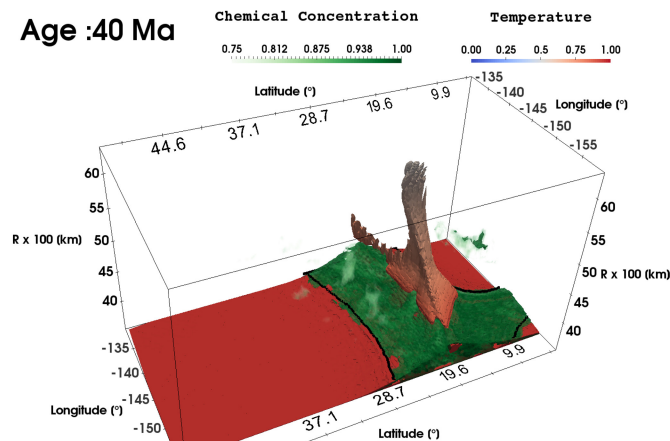
Age :80 Ma



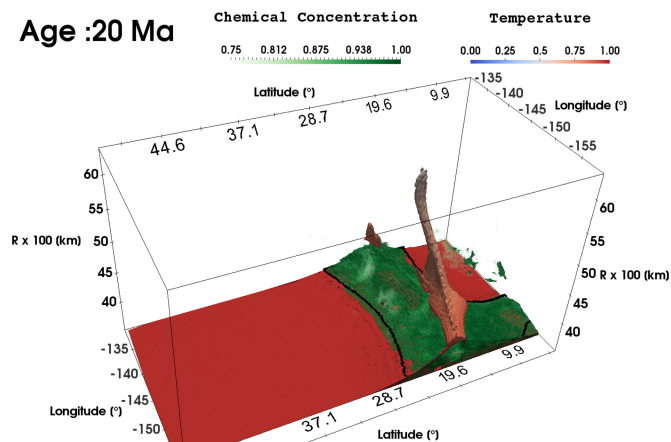
Age :60 Ma



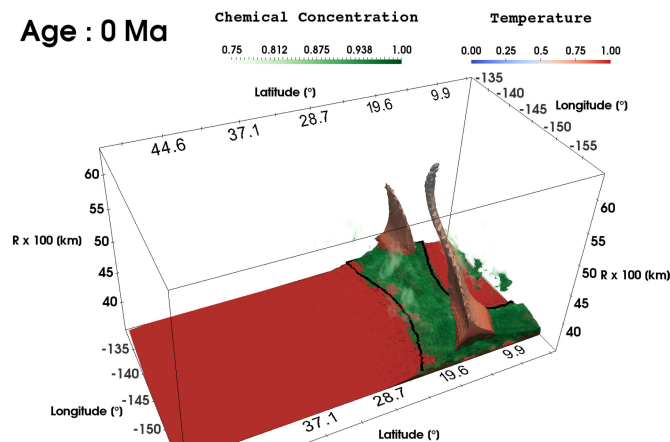
Age :40 Ma



Age :20 Ma



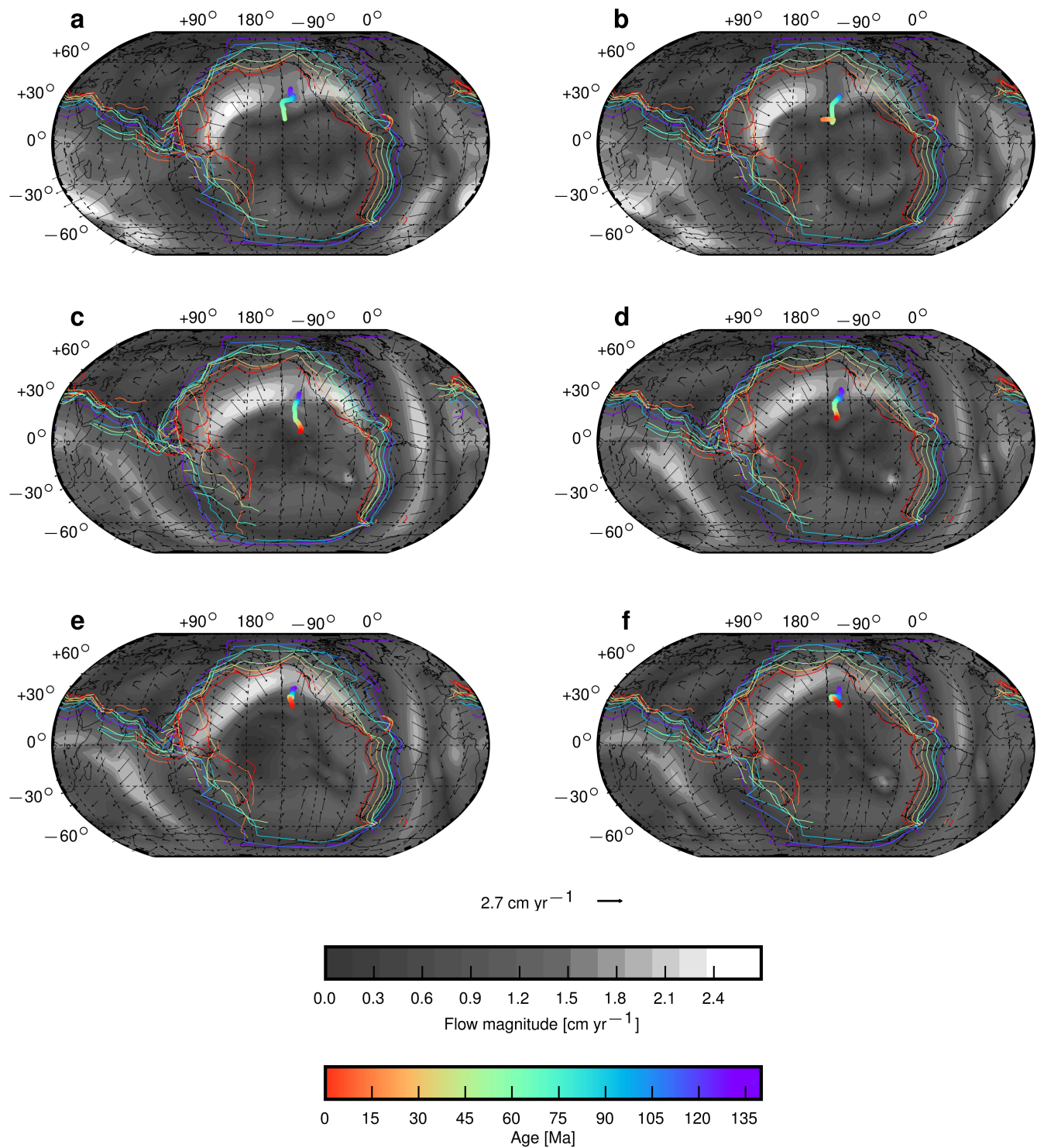
Age : 0 Ma



**Extended Data Figure 3 | Trajectory of model Hawaiian plume.** Three-dimensional (Cartesian projection of spherical geometry) perspectives showing the southward motion and evolution of tilt for model plume corresponding to Hawaii ( $H_m$ ) in case M3 (Extended Data Table 2).

The black contour marks the 75% chemical concentration isosurface 100 km above the CMB. The temperature field above layer averages,  $\delta T$ , is isosurfaced at a value of 0.1 to delineate plume conduits. The top 200 km of the domain is not rendered, in order to avoid visual clutter.





**Extended Data Figure 4 | Inter-model comparisons.** **a**, For case M1 (Extended Data Table 2), the background shading, velocity vectors and subduction zones shown are as described in Fig. 1a. The model plume trajectory for Hawaii ( $H_m$ ) at a depth of 350 km is coloured by age. **b**, For case M2. **c**, For case M4. **d**, For case M5. **e**, For case M6. **f**, For case M7.



Extended Data Table 1 | Physical parameters and constants

Parameter	Symbol	Value	Units
Rayleigh number	$Ra$	$5 \times 10^8$	-
Earth radius	$R_0$	6371	km
Density	$\rho_0$	3930	$\text{kg m}^{-3}$
Thermal expansivity	$\alpha_0$	$1.42 \times 10^{-5}$	$\text{K}^{-1}$
Thermal diffusivity	$\kappa_0$	$1 \times 10^{-6}$	$\text{m}^2 \text{s}^{-1}$
Specific heat capacity	$C_p$	1100	$\text{J kg}^{-1} \text{K}^{-1}$
Gravitational acceleration	$g$	10	$\text{m s}^{-2}$
Surface Temperature	$T_s$	300	K
Dissipation number	$Di$	0.8	-
Reference Viscosity	$\eta_0$	$1 \times 10^{21}$	$\text{Pa s}$
Internal Heating	$H$	100	-

**Extended Data Table 2 | Model cases**

Case	$\Delta \rho_{ch}\%$	APM Model	$\eta_c$	Geometry of Dense Layer
M1	2.125	APM1	1	Uniform
M2	2.25	APM1	1	Uniform
M3	2.5	APM1	1	Uniform
M4	2.5	APM2	1	Uniform
M5	2.5	APM1	1	Domed
M6	2.5	APM1	5	Uniform
M7	3.0	APM1	1	Uniform

$\Delta \rho_{ch}$  is the density contrast of the dense chemical layer above the CMB.

# First North American fossil monkey and early Miocene tropical biotic interchange

Jonathan I. Bloch<sup>1</sup>, Emily D. Woodruff<sup>1,2</sup>, Aaron R. Wood<sup>1,3</sup>, Aldo F. Rincon<sup>1,4</sup>, Arianna R. Harrington<sup>1,2,5</sup>, Gary S. Morgan<sup>6</sup>, David A. Foster<sup>4</sup>, Camilo Montes<sup>7</sup>, Carlos A. Jaramillo<sup>8</sup>, Nathan A. Jud<sup>1</sup>, Douglas S. Jones<sup>1</sup> & Bruce J. MacFadden<sup>1</sup>

New World monkeys (platyrrhines) are a diverse part of modern tropical ecosystems in North and South America, yet their early evolutionary history in the tropics is largely unknown. Molecular divergence estimates suggest that primates arrived in tropical Central America, the southern-most extent of the North American landmass, with several dispersals from South America starting with the emergence of the Isthmus of Panama 3–4 million years ago (Ma)<sup>1</sup>. The complete absence of primate fossils from Central America has, however, limited our understanding of their history in the New World. Here we present the first description of a fossil monkey recovered from the North American landmass, the oldest known crown platyrrhine, from a precisely dated 20.9-Ma layer in the Las Cascadas Formation in the Panama Canal Basin, Panama. This discovery suggests that family-level diversification of extant New World monkeys occurred in the tropics, with new divergence estimates for Cebidae between 22 and 25 Ma, and provides the oldest fossil evidence for mammalian interchange between South and North America. The timing is consistent with recent tectonic reconstructions<sup>2,3</sup> of a relatively narrow Central American Seaway in the early Miocene epoch, coincident with over-water dispersals inferred for many other groups of animals and plants<sup>4</sup>. Discovery of an early Miocene primate in Panama provides evidence for a circum-Caribbean tropical distribution of New World monkeys by this time, with ocean barriers not wholly restricting their northward movements, requiring a complex set of ecological factors to explain their absence in well-sampled similarly aged localities at higher latitudes of North America.

The Miocene epoch (23.8–5.3 Ma) is marked by substantial climatic and ecological changes that had profound effects on terrestrial mammal communities in the New World tropics<sup>5</sup>. Fossils from the tropical lowlands of Central America are rare owing to a lack of relevant rock exposures; however, an important exception can be found in Panama where, since 2009, expansion of the Panama Canal has exposed fossil-bearing rocks of early Miocene age. The lower Miocene Las Cascadas Formation (Fig. 1) represents the oldest fossiliferous continental sequence exposed along the Panama Canal area<sup>3</sup> and includes a diverse fossil mammal assemblage<sup>6–9</sup> that, while compositionally different from that of the younger overlying Centenario Fauna<sup>10</sup>, similarly has almost entirely North American affinities despite its proximity to South America, consistent with a peninsular connection to North America<sup>10,11</sup>. Palaeontological fieldwork in 2012–2013 resulted in the discovery of seven isolated fossil primate teeth that shed new light on the early evolution of crown New World monkeys in the tropics and provide insight into the timing and dynamics of the earliest stages of the Great American Biotic Interchange (GABI) between North and South America<sup>12</sup>.

Primates Linnaeus, 1758

Anthropoidea Mivart, 1864

Platyrrhini Geoffroy, 1812

Cebidae Bonaparte, 1831

*Panamacebus transitus* gen. et sp. nov.

**Etymology.** Generic name combines ‘Panama’ with ‘Cebus’, root taxon for Cebidae. Specific name ‘transit’ (Latin, crossing) refers to its implied early Miocene dispersal between South and North America.

**Holotype.** UF 280128, left upper first molar (M<sup>1</sup>; Fig. 2a, b).

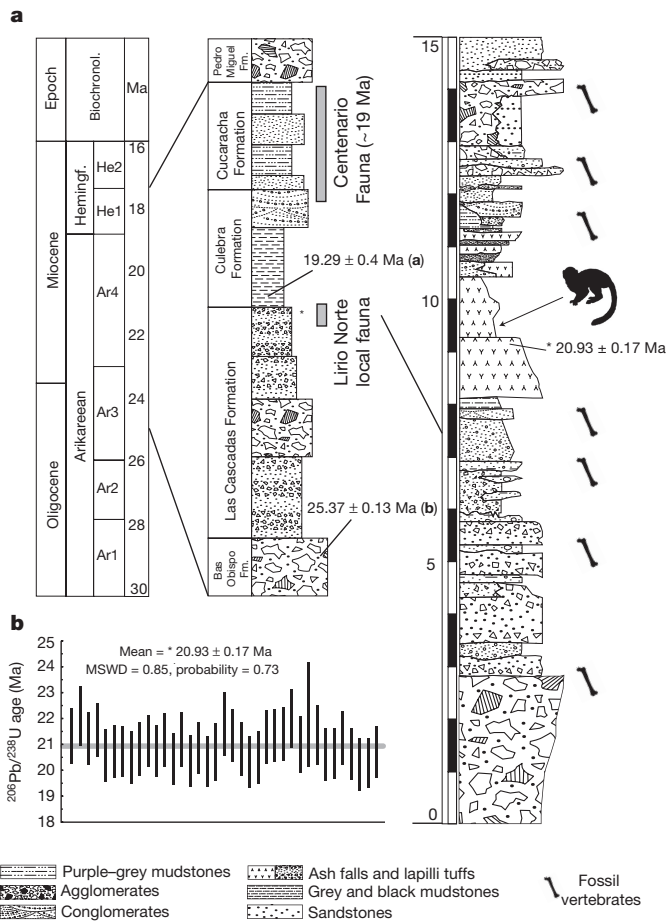
**Referred material.** Left upper second molar (M<sup>2</sup>; UF 281001; Fig. 2a, b), partial left lower first incisor (I<sub>1</sub>; UF 280130), right lower second incisor (I<sub>2</sub>; UF 267048), right lower canine (C<sub>1</sub>; UF 280131), possibly associated left lower second (P<sub>2</sub>; UF 280127) and fourth (P<sub>4</sub>; UF 280129) premolars (Fig. 2e–g).

**Locality.** Lirio Norte (site key YPA-024 in the Florida Museum of Natural History Vertebrate Paleontology Collection), Panama Canal area, Panama, Central America (Extended Data Fig. 1).

**Age and horizon.** Primate fossils were collected from a single horizon of sub-aerially exposed ash fall deposits in the upper part of the Las Cascadas Formation (Fig. 1a). Magmatic zircons from a non-subaerially exposed andesitic tuff, ~0.25 m below the primate-bearing horizon (Extended Data Fig. 2), yield an age of 20.93 ± 0.17 Ma (2σ) (Fig. 1b and Supplementary Table 1), interpreted to be the eruption age of the tuff and a close approximation of the absolute age of the primate-bearing horizon. The mammalian assemblage (Supplementary Table 6) recovered from this horizon is consistent with what is known from the late Arikareean Ar4 faunal zone (22.8–19.05 Ma; refs 10, 13) North American Land Mammal Age (NALMA) at higher latitudes<sup>13,14</sup> (Extended Data Fig. 3). See also Supplementary Information (results section).

**Diagnosis.** Medium-sized cebid platyrrhine (~2.7 kg) that differs from all other cebines in having lower incisors (I<sub>1</sub>, I<sub>2</sub>) that are higher crowned with an incomplete lingual cingulum, and a first upper molar (M<sup>1</sup>) that is considerably larger than the second upper molar (M<sup>2</sup>). Further differs from most cebines: (except *Aotus*) in having an I<sub>1</sub> cross-sectional area only somewhat smaller than that of I<sub>2</sub>; (except *Cebus*) in having somewhat inflated lower premolars (P<sub>2</sub>, P<sub>4</sub>), and a P<sub>4</sub> talonid and trigonid of similar length; and (except *Saimiri* and *Neosaimiri*) in having M<sup>1,2</sup> with a short and strong anterior cingulum. Further differs: from *Cebus* and *Acrecebus* in having M<sup>1,2</sup> with a mesiobuccally oriented prehypocrista, lacking a metaconule, a postprotocrista that is continuous with a hypometacrista that reaches to the base of the metacone, a sharp and distinct hypometacrista, and lack of lingual inflation of the protocone; from *Cebus* in having a P<sub>4</sub> with a smaller hypoconid; and from *Saimiri* in lacking a pericone on M<sup>1,2</sup>. Differs from all callitrichines in having a discrete

<sup>1</sup>Florida Museum of Natural History, University of Florida, Gainesville, Florida 32611-7800, USA. <sup>2</sup>Department of Biology, University of Florida, Gainesville, Florida 32611-7800, USA. <sup>3</sup>Department of Geological and Atmospheric Sciences, Iowa State University, Ames, Iowa 50011-1027, USA. <sup>4</sup>Department of Geological Sciences, University of Florida, Gainesville, Florida 32611-7800, USA. <sup>5</sup>Department of Evolutionary Anthropology, Duke University, Durham, North Carolina 27708-9976, USA. <sup>6</sup>New Mexico Museum of Natural History and Science, Albuquerque, New Mexico 87104, USA. <sup>7</sup>Geociencias, Universidad de los Andes, Calle 1A # 18A-10, Edificio IP, Bogotá DC 111711, Colombia. <sup>8</sup>Smithsonian Tropical Research Institute, Box 0843-03092, Balboa, Ancon, Republic of Panama.



**Figure 1 | Stratigraphy of the primate-bearing locality (YPA-024) in central Panama.** **a**, Measured stratigraphic section (in metres) in the Las Cascadas Formation showing the positions of the dated rock sample (asterisk) and of the *Panamacebus* fossils as a silhouette of a monkey (right), correlated to a schematic stratigraphic position of the Lirio Norte Local Fauna and the Centenario Fauna with previously published radiometric dates (**a**, ref. 3; **b**, ref. 28) indicated (centre), and North American land mammal faunal zonation (left). Biochronol., Biochronology; Fm., formation; Hemingf., Hemingfordian North American Land Mammal Age; MSWD, mean square of weighted deviates. **b**, Results from U–Pb isotopic analyses plotted as a weighted mean of the  $^{206}\text{Pb}/^{238}\text{U}$  ages from 37 zircons (see Methods and Supplementary Information).

entoconid on  $P_4$ , and  $M^{1,2}$  with a large hypocone and lacking a buccal cingulum; from callitrichines (but also *Aotus*) in having an  $M^2$  with a buccally expanded paracone; and from callitrichines (but also *Saimiri* and *Neosaimiri*) in having  $M^{1,2}$  with a well-developed prehypocrista that extends to the postprotocrista and encloses the talon lingually. For additional description and metrics see Extended Data Figs 4–7 and Supplementary Information (results section).

*Panamacebus* is similar to other platyrrhines in having a single-rooted  $P_4$  rather than double-rooted as in catarrhines. Among cebids, the dental morphology is most similar to that of fossil cebines, including middle Miocene *Neosaimiri* from Colombia (Fig. 2 and Extended Data Figs 4–7). For additional comparisons see Supplementary Information (results section) and Supplementary Figs 2 and 3.

Until recently, the oldest evidence for the arrival of primates in South America from the Old World was from the late Oligocene epoch (26 Ma) of Bolivia<sup>15</sup>. The recent discovery of anthropoids from eastern Peru<sup>16</sup> suggests an earlier arrival into tropical South America, perhaps during the late Eocene epoch (~37–34 Ma), although the date is uncertain<sup>17</sup>. Dispersals from Africa or Asia through Antarctica or North America have been suggested<sup>18</sup>, but ‘rafting’ on floating islands from Africa across the Atlantic is considered the most likely mechanism

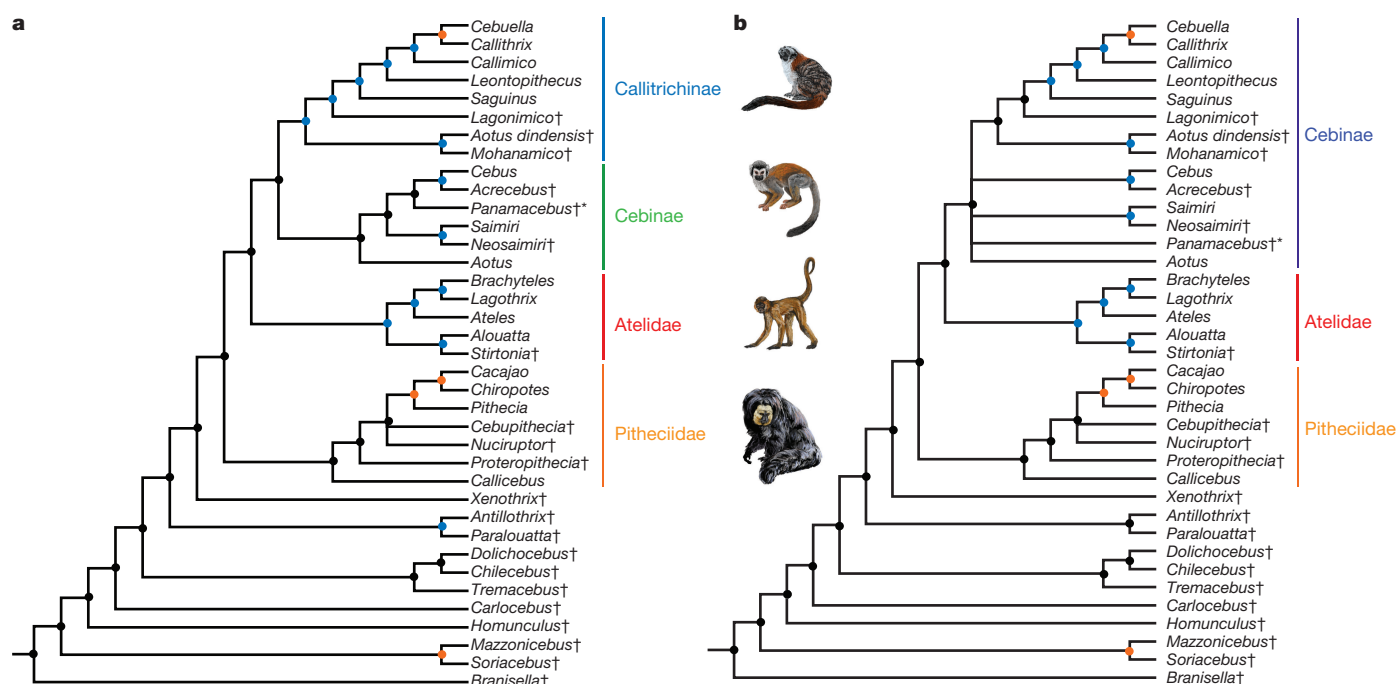


**Figure 2 | Comparison of *Panamacebus* with middle Miocene cebid *Neosaimiri fieldsi* from La Venta, Colombia.** **a**, **b**, Occlusal (**a**) and lingual (**b**) views of *P. transitus* left  $M^1$  (UF 280128: <http://dx.doi.org/10.17602/M2/M8531>) and  $M^2$  (UF 281001: <http://dx.doi.org/10.17602/M2/M8550>). **c**, **d**, Occlusal (**c**) and lingual (**d**) views of *N. fieldsi* right  $M^1$  (IGM-KU 89008, mirrored: <http://dx.doi.org/10.17602/M2/M8541>) and  $M^2$  (IGM-KU 89018, mirrored: <http://dx.doi.org/10.17602/M2/M8543>). **e**–**g**, Occlusal (**e**), lingual (**f**) and buccal (**g**) views of *P. transitus* partial left  $I_1$  (UF 280130: <http://dx.doi.org/10.17602/M2/M8400>), right  $I_2$  (UF 267048, mirrored: <http://dx.doi.org/10.17602/M2/M8395>), right  $C_1$  (UF 280131, mirrored: <http://dx.doi.org/10.17602/M2/M8401>), left  $P_2$  (UF 280127: <http://dx.doi.org/10.17602/M2/M8397>) and left  $P_4$  (UF 280129: <http://dx.doi.org/10.17602/M2/M8399>). **h**–**j**, Occlusal (**h**), lingual (**i**) and buccal (**j**) views of *N. fieldsi* left dentary with  $I_2$ – $M_2$  (UCMP 39205: <http://dx.doi.org/10.17602/M2/M1879>, <http://dx.doi.org/10.17602/M2/M1880>). Images generated from microCT scan data (see Methods and links associated with specimen numbers). Scale bars: 1 mm (**a**–**d**); 5 mm (**e**–**j**).

for their arrival into South America<sup>19</sup>. Results from phylogenetic analyses of new morphological data coded into a previously published matrix<sup>1</sup> vary with the use of different molecular constraints<sup>1,20</sup>, but both support *Panamacebus* within crown Platyrrhini, specifically within Cebidae (Fig. 3; Supplementary Information (results section) and Supplementary Figs 4–10), suggesting a northward early Miocene dispersal across the Central American Seaway (CAS) from South to North America, perhaps associated with the onset of extensive development of terrestrial landscapes in Central America as a consequence of the initial collision with South America<sup>2,3</sup> (Fig. 4 and Extended Data Fig. 8).

Using our most resolved tree (Fig. 3a), the minimum age for a split between Callitrichinae and Cebinae can be calibrated with the radioisotopic date (20.77–21.90 Ma, posterior probability (PP): 1.0). The results of our divergence dating analysis (Extended Data Figs 9, 10 and Supplementary Information (results section)) using this date as a prior (Supplementary Table 2), considerably narrow previously reported range estimates (16.07–23.5 Ma; ref. 20; and 15.66–24.03 Ma; ref. 21). Our inferred divergence of Cebidae from Atelidae (21.84–24.93 Ma, PP: 0.99) also greatly reduces the interval estimated in a previous study (18.14–26.11 Ma; ref. 20). Our divergence estimates for other primate clades are generally congruent with previous studies (Supplementary Information (results section)). *Panamacebus* does not seem to share a





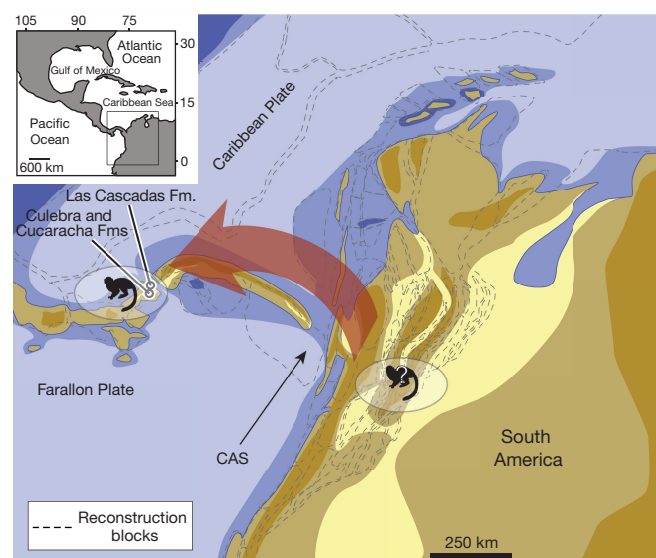
**Figure 3 | Results from phylogenetic analyses showing *Panamacebus* within crown Platyrrhini. a**, Strict consensus of two most parsimonious trees analysed using molecular constraint 1 (ref. 1). **b**, Strict consensus of four most parsimonious trees analysed using molecular constraint 2 (ref. 20). Circles at nodes correspond to bootstrap support values: orange, greater than 85; blue, 50–85; black, less than 50. Daggers indicate fossil (extinct) taxa. Asterisk indicates *P. transitus*.

close relationship with fossil primates from the Quaternary period of Cuba, Jamaica and Hispaniola (Fig. 3a, b), although they probably also dispersed from tropical South America into the Greater Antilles by the early Miocene<sup>1,22</sup>. This result suggests two unrelated migrations from South America to both tropical North America (a crown platyrrhine) and the Caribbean islands (probable stem platyrrhines) during the Oligocene/early Miocene epochs.

While attribution of fossils to crown versus stem platyrrhines has been the subject of substantial ongoing debate<sup>23</sup>, our results support an exclusively tropical crown platyrrhine radiation with only the more distantly related (stem) platyrrhines distributed into the higher latitudes of South America<sup>1</sup>. However, if the alternative view proves to be correct, that primates of similar age in Patagonia are also crown platyrrhines (the long-lineage hypothesis)<sup>23</sup>, *Panamacebus* would still be at least equal in age to the oldest disputed crown platyrrhine. Coupled with fossil evidence for primates in the lower Miocene of the Greater Antilles<sup>24</sup>, this new record in Central America shows that primates were dispersing northward out of South America with a previously unrecognized early Miocene circum-Caribbean distribution. The primate record is consistent with new tectonic and palaeogeographic reconstructions<sup>2,3</sup> of a relatively narrow CAS in the early Miocene and corresponding dispersals inferred on the basis of molecular and fossil data for many terrestrial organisms, including amphibians, reptiles, freshwater fishes, insects and plants<sup>4</sup>. In this context, it is perhaps surprising to not also find fossil evidence of caviomorph rodents and sloths in the early Miocene of Panama since they are found in a slightly younger locality that includes evidence of a primate in the Greater Antilles<sup>24</sup>. Absence of primates from the overlying Centenario Fauna<sup>10</sup>, which is ~2 million years (Myr) younger than the Lirio Norte Local Fauna, might suggest a single waif dispersal event of a short-lived population. Alternatively, with increased collecting efforts the fossil record may yet provide evidence for other earlier mammal dispersals out of South America, and/or primates having a longer history in tropical North America than is currently known.

Prior to this discovery, the oldest fossil evidence for mammalian dispersal from South to North America is 8.5–9 Ma mylodontid and

megalonchid sloths that, along with later immigrants associated with different pulses of the GABI, would have necessarily had to move through the tropics, but also quickly dispersed to higher latitudes. By contrast, New World monkeys clearly crossed into the tropical lowlands of Central America at least once by 21 Ma, but there is no record of them in localities of similar age at higher northern latitudes. This is especially notable in the Gulf Coastal Plain, where most other mammals



**Figure 4 | Palaeogeographic reconstruction showing hypothetical dispersal route of *Panamacebus* across the CAS in the early Miocene.** Yellow and ochre colours indicate subaerial environments, blue colours indicate marine environments (dark, coastal and platform; light, abyssal). Criteria used to arrive at this reconstruction include regional tectonic reconstructions, local and regional palaeomagnetic data, and regional strain markers and piercing points (see Extended Data Fig. 8, Methods, and Supplementary Methods). Fm., formation; Fms, formations.

share close affinities with those found in Panama<sup>6–9</sup>. Absence of early Miocene New World monkeys at higher northern latitudes could be explained by the limited extent of suitable tropical forest habitats, much like today. However, their distribution in South America, including very high latitudes of Patagonia, at localities of very similar age<sup>25</sup>, introduces a potential paradox in which primates would be limited to tropical forests in North but not South America<sup>26</sup>. A possible resolution is found in the taxonomic composition and historical biogeography of the forests themselves. Early Miocene forests of tropical South America have a shared Gondwanan history with those at higher southern latitudes, including Patagonia, and southern Central America (Costa Rica and Panama), which are dominated by South American-derived tropical rainforest taxa<sup>27</sup>. Northern tropical Central American forests, however, have predominantly Laurasian affinities in the early Miocene (Supplementary Tables 3 and 7 and Supplementary Information (results section)). Thus, dispersal of New World monkeys further northward in the early Miocene was probably limited more by their niche conservatism and a boundary between forests with very different evolutionary histories than by differences in climate or the existence of major geographical barriers. The New World tropics may have acted as a holding pen for some tropical mammals for at least 12 Myr before xenarthrans first appeared at higher latitudes of North America<sup>12</sup>, suggesting that the distribution of South American-derived tropical forests played an important part in the early dynamics of the GABI.

**Online Content** Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 4 November 2015; accepted 09 February 2016.

Published online 20 April 2016.

- Kay, R. F. Biogeography in deep time—what do phylogenetics, geology, and paleoclimate tell us about early platyrrhine evolution? *Mol. Phylogenet. Evol.* **82**, 358–374 (2015).
- Farris, D. W. *et al.* Fracturing of the Panamanian Isthmus during initial collision with South America. *Geology* **39**, 1007–1010 (2011).
- Montes, C. *et al.* Evidence for middle Eocene and younger land emergence in central Panama: implications for Isthmus closure. *Geol. Soc. Am. Bull.* **124**, 780–799 (2012).
- Bacon, C. D. *et al.* Biological evidence supports an early and complex emergence of the Isthmus of Panama. *Proc. Natl Acad. Sci. USA* **112**, 6110–6115 (2015).
- MacFadden, B. J. Extinct mammalian biodiversity of the ancient New World tropics. *Trends Ecol. Evol.* **21**, 157–165 (2006).
- Rincon, A. F., Bloch, J. I., Suarez, C., MacFadden, B. J. & Jaramillo, C. A. New floridatragulines (Mammalia, Camelidae) from the early Miocene Las Cascadas Formation, Panama. *J. Vertebr. Paleontol.* **32**, 456–475 (2012).
- Rincon, A. F., Bloch, J. I., Macfadden, B. J. & Jaramillo, C. A. First Central American record of Anthracotheriidae (Mammalia, Bothriodontinae) from the early Miocene of Panama. *J. Vertebr. Paleontol.* **33**, 421–433 (2013).
- Rincon, A. F., Bloch, J. I., Macfadden, B. J. & Jaramillo, C. A. New early Miocene protoceratids (Mammalia, Artiodactyla) from Panama. *J. Vertebr. Paleontol.* <http://dx.doi.org/10.1080/02724634.2015.970688> (2015).
- Wood, A. R. & Ridgwell, N. M. The first Central American chalicotheres (Mammalia, Perissodactyla) and the paleobiogeographic implications for small-bodied schizotheriines. *J. Vertebr. Paleontol.* **35**, e923893 (2015).
- MacFadden, B. J. *et al.* Temporal calibration and biochronology of the Cenarian Fauna, early Miocene of Panama. *J. Geol.* **122**, 113–135 (2014).
- Kirby, M. X., Jones, D. S. & MacFadden, B. J. Lower Miocene stratigraphy along the Panama Canal and its bearing on the Central American Peninsula. *PLoS One* **3**, e2791 (2008).
- Woodburne, M. O. The Great American Biotic Interchange: dispersals, tectonics, climate, sea level and holding pens. *J. Mamm. Evol.* **17**, 245–264 (2010).
- Albright, L. B., III *et al.* Revised chronostratigraphy and biostratigraphy of the John Day Formation (Turtle Cove and Kimberly Members), Oregon, with implications for updated calibration of the Arikarean North American land mammal age. *J. Geol.* **116**, 211–237 (2008).
- Tedford, R. H. *et al.* in *Late Cretaceous and Cenozoic Mammals of North America: Biostratigraphy and Geochronology* 169–231 (Columbia Univ. Press, 2004).
- MacFadden, B. J. Chronology of Cenozoic primate localities in South America. *J. Hum. Evol.* **19**, 7–21 (1990).
- Bond, M. *et al.* Eocene primates of South America and the African origins of New World monkeys. *Nature* **520**, 538–541 (2015).
- Kay, R. F. New World monkey origins. *Science* **347**, 1068–1069 (2015).
- Jameson Kiesling, N. M., Yi, S. V., Xu, K., Gianluca Sperone, F. & Wildman, D. E. The tempo and mode of New World monkey evolution and biogeography in the context of phylogenomic analysis. *Mol. Phylogenet. Evol.* **82**, 386–399 (2015).
- Houle, A. The origin of platyrrhines: an evaluation of the Antarctic scenario and the floating island model. *Am. J. Phys. Anthropol.* **109**, 541–559 (1999).
- Springer, M. S. *et al.* Macroevolutionary dynamics and historical biogeography of primate diversification inferred from a species supermatrix. *PLoS ONE* **7**, e49521 (2012).
- Perelman, P. *et al.* A molecular phylogeny of living primates. *PLoS Genet.* **7**, e1001342 (2011).
- Cooke, S. B., Rosenberger, A. L. & Turvey, S. An extinct monkey from Haiti and the origins of the Greater Antillean primates. *Proc. Natl Acad. Sci. USA* **108**, 2699–2704 (2011).
- Rosenberger, A. L. Platyrrhines, PAUP, parallelism, and the Long Lineage Hypothesis: a reply to Kay *et al.* (2008). *J. Hum. Evol.* **59**, 214–217 (2010).
- MacPhee, R. D., Iturralde-Vinent, M. A. & Gaffney, E. S. Domo de Zaza, an early Miocene vertebrate locality in south-central Cuba, with notes on the tectonic evolution of Puerto Rico and the Mona Passage. *Am. Mus. Novit.* **3394**, 1–42 (2003).
- Dunn, R. E. *et al.* A new chronology for middle Eocene–early Miocene South American Land Mammal Ages. *Geol. Soc. Am. Bull.* **125**, 539–555 (2013).
- Rosenberger, A., Tejedor, M., Cooke, S. & Pekar, S. in *South American Primates Developments in Primatology: Progress and Prospects* (eds Garber, P. A., Estrada, A., Bicca-Marques, J. C., Heymann, E. W. & Strier, K. B.) Ch. 4, 69–113 (Springer, 2009).
- Jaramillo, C. *et al.* *Palynological Record of the Last 20 Million Years in Panama*. Ch. 8, 134–251 (Missouri Botanical Garden Press, 2014).
- Rooney, T. O., Franceschi, P. & Hall, C. M. Water-saturated magmas in the Panama Canal region: a precursor to adakite-like magma generation? *Contrib. Mineral. Petrol.* **161**, 373–388 (2011).

Supplementary Information is available in the online version of the paper.

**Acknowledgements** We thank M. Silcox, D. Boyer, G. Gunnell, S. Chester, P. Morse, E. Sargis, D. Steadman, E. Kowalski, Z. Randall, A. Rosenberger, J. Krigbaum and D. Daegling for advice and discussion, R. Kay and L. Marivaux for reviews that significantly improved the quality of the manuscript, D. Byerley and G. Maccracken for finding primate fossils in Panama, M. Drouillard for assistance with geochronology laboratory and fieldwork, R. Kay, P. Holroyd and D. Reed for access to comparative specimens, J. Bourque for fossil preparation, D. Byerley for artwork associated with Fig. 3, D. Boyer for facilitating access to the Duke University SMIF microCT facility, and D. Boyer, G. Yapuncich and J. Thostenson for help acquiring and processing microCT scan data (funded in part by National Science Foundation (NSF) BCS 1304045 to D. Boyer and E. St Clair, and BCS 0851272 to R. Kay). We thank O. Moskalenko, M. Gitzendanner and D. Reed for assistance with the high-performance computing resources at the University of Florida. We thank the Autoridad del Canal de Panama (ACP) and the Ministerio de Comercio e Industria (MICI) for logistical support and access to the Panama Canal Zone. Part of this manuscript was written when J.I.B. was supported as an Edward P. Bass Distinguished Visiting Environmental Scholar in the Yale Institute for Biospheric Studies (YIBS). The NSF (PIRE project 0966884), Smithsonian Tropical Research Institute Paleobiology Fund, and the Florida Museum of Natural History funded this research. This is University of Florida Contribution to Paleobiology 782.

**Author Contributions** J.I.B., A.R.W., E.D.W. and G.S.M. contributed to project planning. J.I.B., A.R.W. and A.R.H. contributed to systematic palaeontology and microCT scans. D.A.F. and A.F.R. contributed to radioisotopic analyses and stratigraphy. B.J.M., A.F.R., G.S.M., A.R.W. and J.I.B. contributed to biochronological analysis. E.D.W., J.I.B. and A.R.W. contributed to phylogenetic analyses. E.D.W. performed divergence dating analyses. C.M. and C.A.J. contributed to palaeogeographic analysis. N.A.J., J.I.B. and C.A.J. contributed to the pollen summary. A.R.W., A.F.R., J.I.B., G.S.M., E.D.W. and D.S.J. contributed to fieldwork. J.I.B., B.J.M., G.S.M., C.A.J. and D.S.J. contributed to financial support. All authors contributed to manuscript and figure preparation.

**Author Information** The LSIDs for *Panamacebus* (genus), urn:lsid:zoobank.org:act:C3F8967-EE79-47B8-8A98-2C6D2C7557CA, and for *Panamacebus transitus* (species), urn:lsid:zoobank.org:act:3E01F3F2-B1F8-433E-B110-9AD5AAF3DB99, have been deposited in ZooBank. Reprints and permissions information is available at [www.nature.com/reprints](http://www.nature.com/reprints). The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to J.I.B. (jbloch@flmnh.ufl.edu).

## METHODS

**Geochronology.** Sample MD11 was collected from an ash layer that forms the upper-most part of a several-metre thick, welded andesitic tuff. The tuff is exposed within the conformable section of the Las Cascadas Formation approximately 0.25 m below the mammal fossil-bearing unit (Fig. 1). Zircons were separated from two 10-kg sample aggregates using conventional magnetic and density concentration methods. Clear to light pink, euhedral zircons about 100 µm in length were mounted in epoxy and ground to reveal internal surfaces.

U–Pb isotopic analyses were conducted on 20-µm spots on zircon grains using a Nu-Plasma multicollector plasma source mass spectrometer coupled to a New Wave 213-nm laser (LA-MC-ICP-MS). The data were acquired using the Nu-Instruments Time Resolved Analysis software. Data calibration and drift were based on multiple analyses of the reference zircon FC-1 (ref. 29) repeated between every 10 ablations of the unknowns. Concordant U–Pb data were yielded by 37 zircons (Supplementary Table 1). A weighted mean of the  $^{206}\text{Pb}/^{238}\text{U}$  ages gives  $20.93 \pm 0.17$  Ma ( $2\sigma$ ) (Fig. 1b), which we interpret to be the eruption age of the andesitic tuff and the absolute age of the Las Cascadas mammal fossil horizon (Supplementary Information (results section)). These data are consistent with a  $^{206}\text{Pb}/^{238}\text{U}$  age of  $19.3 \pm 0.4$  Ma for a welded tuff unit within the younger Culebra Formation<sup>30</sup>.

**Three-dimensional data acquisition.** Fossil platyrrhine specimens (UF 267048, 28001, 280127–280131; UCMP 38762, 38989, 39205) and casts (IGM-KU 89008, 89011, 89018, 89019, 89021, 89086, 89092, 89104, 90016) were scanned at the Duke University Shared Materials Instrumentation facility in Durham, North Carolina, using a Nikon XTH 225 ST MicroCT scanner (Supplementary Table 4). The resulting scans were reconstructed into tiff stacks and were imported into Avizo 7 (Visualization Sciences Group) for surface visualization (Fig. 2, Extended Data Figs 4–7 and Supplementary Figs 2, 3) and manipulation. Data were distributed and shared using Duke University's three-dimensional data archive (<http://www.morphosource.org>). DOI links to three-dimensional data are provided in the figure captions.

**Comparative morphology.** Comparisons were done within the morphological framework outlined and discussed previously<sup>1</sup>, with reference to the literature for fossil taxa as discussed in Supplementary Methods (see also table 1 in ref. 1 and references therein). Additional comparisons were made to specimens and casts of fossil and extant platyrrhines. Three-dimensional images of relevant fossil platyrrhines were reconstructed from microCT scans (see earlier).

**Phylogenetic analysis of morphological data.** We conducted three phylogenetic analyses using maximum parsimony to examine the phylogenetic position of *Panamacebus*. The first analysis was conducted using the same parameters as the original analysis<sup>1</sup> with 177 ordered characters and all characters given a weight other than one (Supplementary Data 1). In the second analysis we used a different constraint tree derived from a recent molecular supermatrix<sup>20,21</sup> (re-analyses of these data used for the constraint tree are described in Supplementary Methods). The third analysis was performed without enforcing a constraint tree. Support for the topology of the resulting trees was determined

by bootstrapping (see Supplementary Methods for detailed descriptions of phylogenetic analyses).

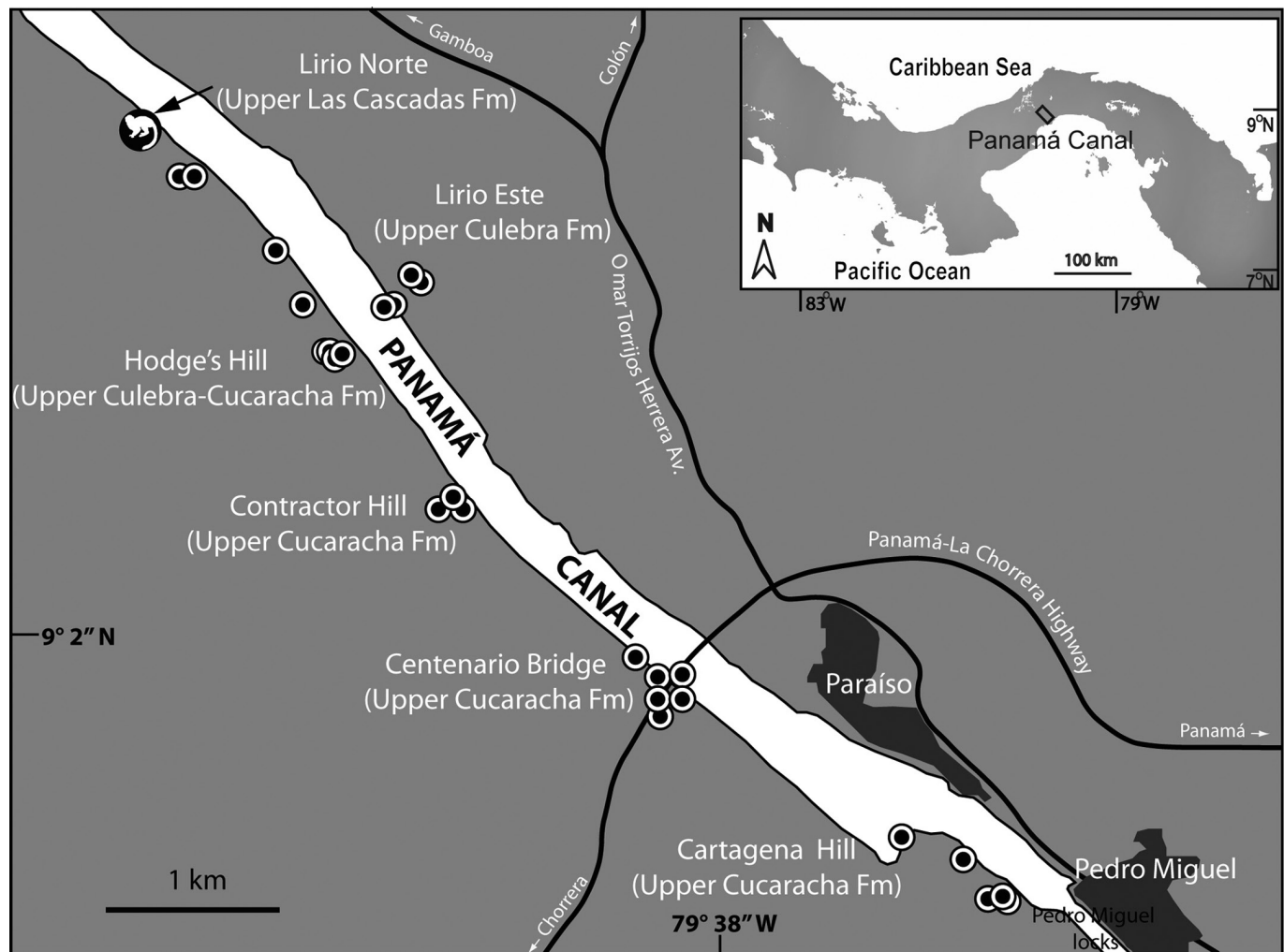
**Divergence dating recalibration analyses.** Divergence dating analyses were conducted to determine the effect of a new minimum age for the divergence of Callitrichinae and Cebinae based on the radioisotopic date obtained for the horizon where the *Panamacebus* fossils were discovered. We took a conservative approach and calibrated the divergence of Cebinae and Callitrichinae as opposed to calibrating a node within Cebinae as suggested by the placement of *Panamacebus* in the phylogeny using the first molecular constraint (Fig. 3a). Using a reduced molecular data set derived from ref. 20, we conducted two separate Bayesian Markov chain Monte Carlo (MCMC) analyses (see Supplementary Methods and Supplementary Data 3). Our analyses employed a lognormal relaxed molecular clock and a general time reversible (GTR) model with a gamma distribution, using four rate categories and estimated base frequencies. We set 15 calibration points (Supplementary Table 2 and Supplementary Data 4); 14 nodes were calibrated as described previously<sup>20</sup> and a 15th node was calibrated to the date corresponding to the age of *Panamacebus* ( $20.93 \pm 0.17$  Ma). A random starting tree was used and analyses were conducted with an MCMC chain length of 200,000,000 states sampled every 10,000 states (for detailed description of divergence dating analyses see Supplementary Methods).

**Palaeogeographic reconstruction.** Figure 4 and Extended Data Fig. 8 show a reconstruction built using fault-bound tectonic blocks restored to a late Oligocene–early Miocene palinspastic palaeogeographic position. Criteria used to arrive at this reconstruction include: (1) regional tectonic reconstructions; (2) local and regional palaeomagnetic data; and (3) regional strain markers and piercing points. Published stratigraphic columns with age constraints spanning the late Oligocene–early Miocene were placed in their corresponding palaeogeographic position. See Supplementary Methods for detailed methods and references.

**Palaeobotanical analysis of Miocene forests.** We compiled pollen and spore occurrence data from nine different formations of Oligocene to middle Miocene age (Supplementary Table 3). The samples come from Florida, Puerto Rico, southern Mexico, Costa Rica and Panama; the number of samples per formation ranges from 2 to 52 (Supplementary Table 7). We assigned the plant families and genera to coarse biogeographic categories on the basis of modern species distributions and fossil evidence. The categories are Gondwana–Amazonian, Gondwana–northern Andean, Gondwana–southern Andean, Laurasian, or unassigned. Next, we calculated the percentage contribution of the biogeographic regions to the species (morphotype) richness in each fossil flora. Taxa that were not assigned to a biogeographic region were excluded from calculations of biogeographic affinity as described previously<sup>27</sup>. See Supplementary Methods for detailed references and discussion.

29. Black, L. P. *et al.* The application of SHRIMP to Phanerozoic geochronology; a critical appraisal of four zircon standards. *Chem. Geol.* **200**, 171–188 (2003).

30. Montes, C. *et al.* Arc-continent collision and orocline formation: closing of the Central American seaway. *J. Geophys. Res.* **117**, B4 (2012).

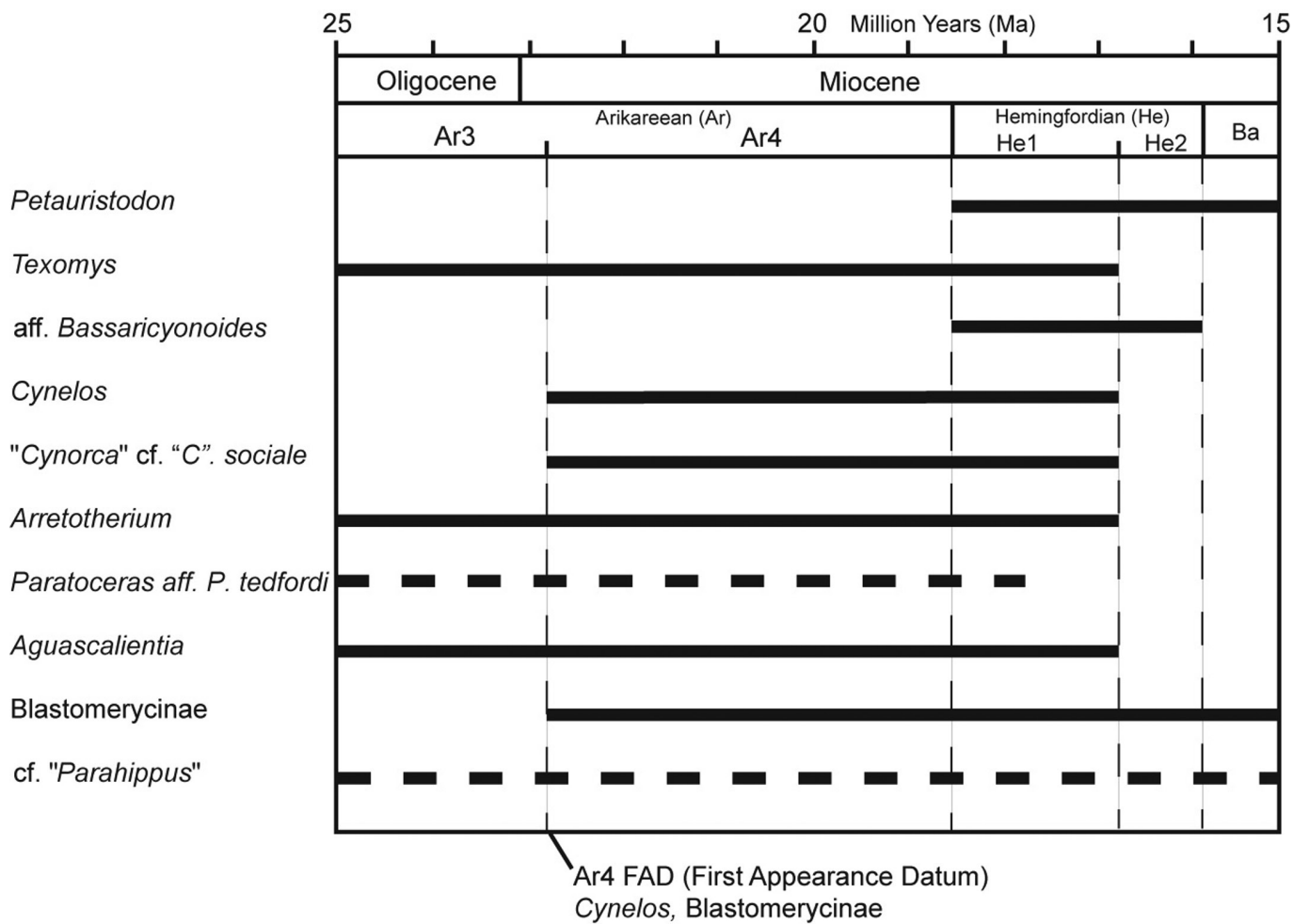


**Extended Data Figure 1 | Map of the southern part of Panama Canal (Gaillard Cut).** Black circles mark the position of the Lirio Norte Local Fauna locality (YPA024) and other specific terrestrial vertebrate collecting sites (Centenario Fauna) in the area (modified from ref. 10).

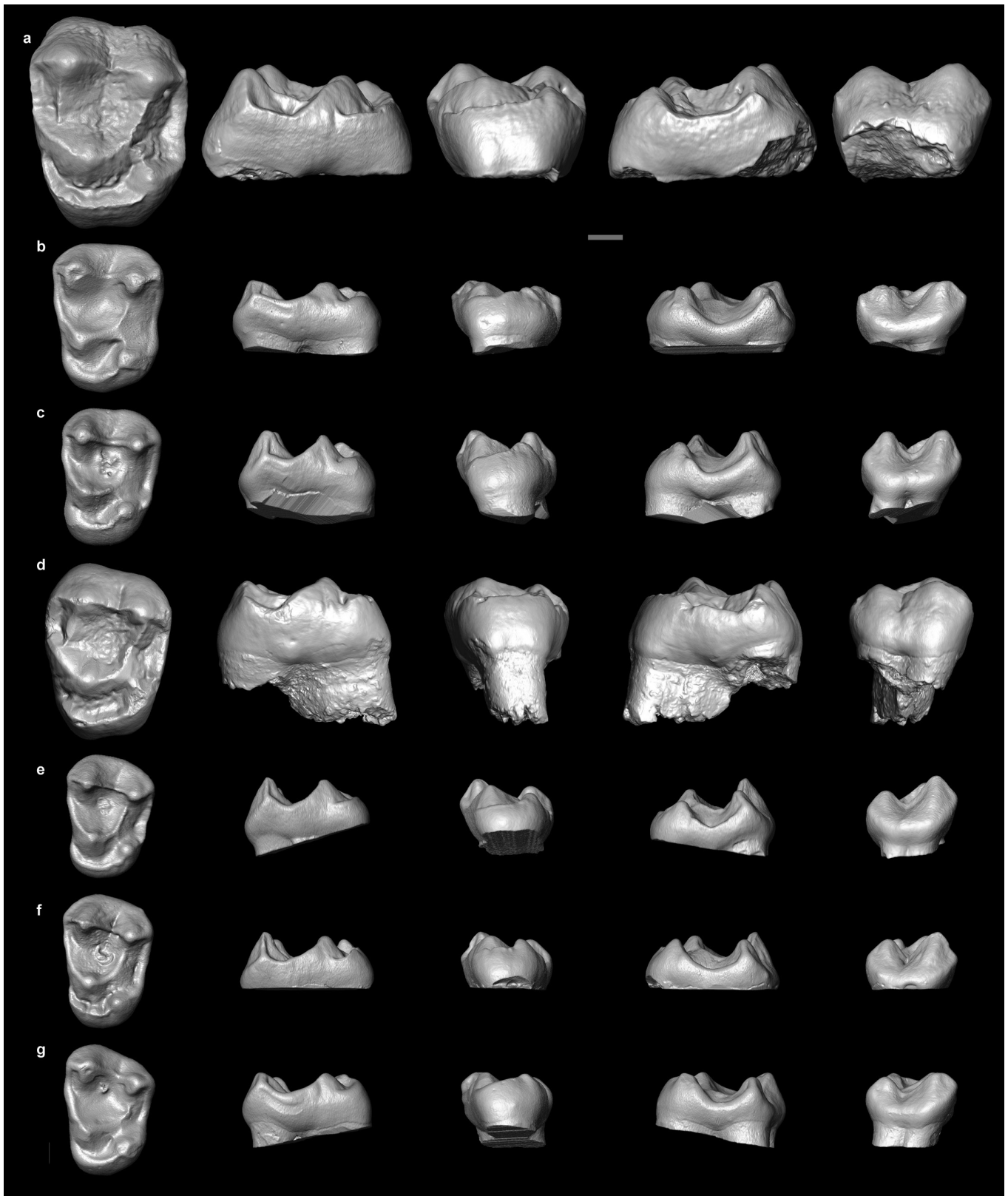




**Extended Data Figure 2 | Photograph of the northern wall (south-facing) in the Lirio Norte Local Fauna locality (YPA024).** Dated rock sample MD11 was collected from an ash layer that forms the upper-most part of a several-metre thick, welded andesitic tuff. The tuff is exposed within the conformable section of the Las Cascadas Formation approximately 0.25 m below, and in close proximity to, the mammal fossil-bearing unit that includes the *Panamacebus* fossils.



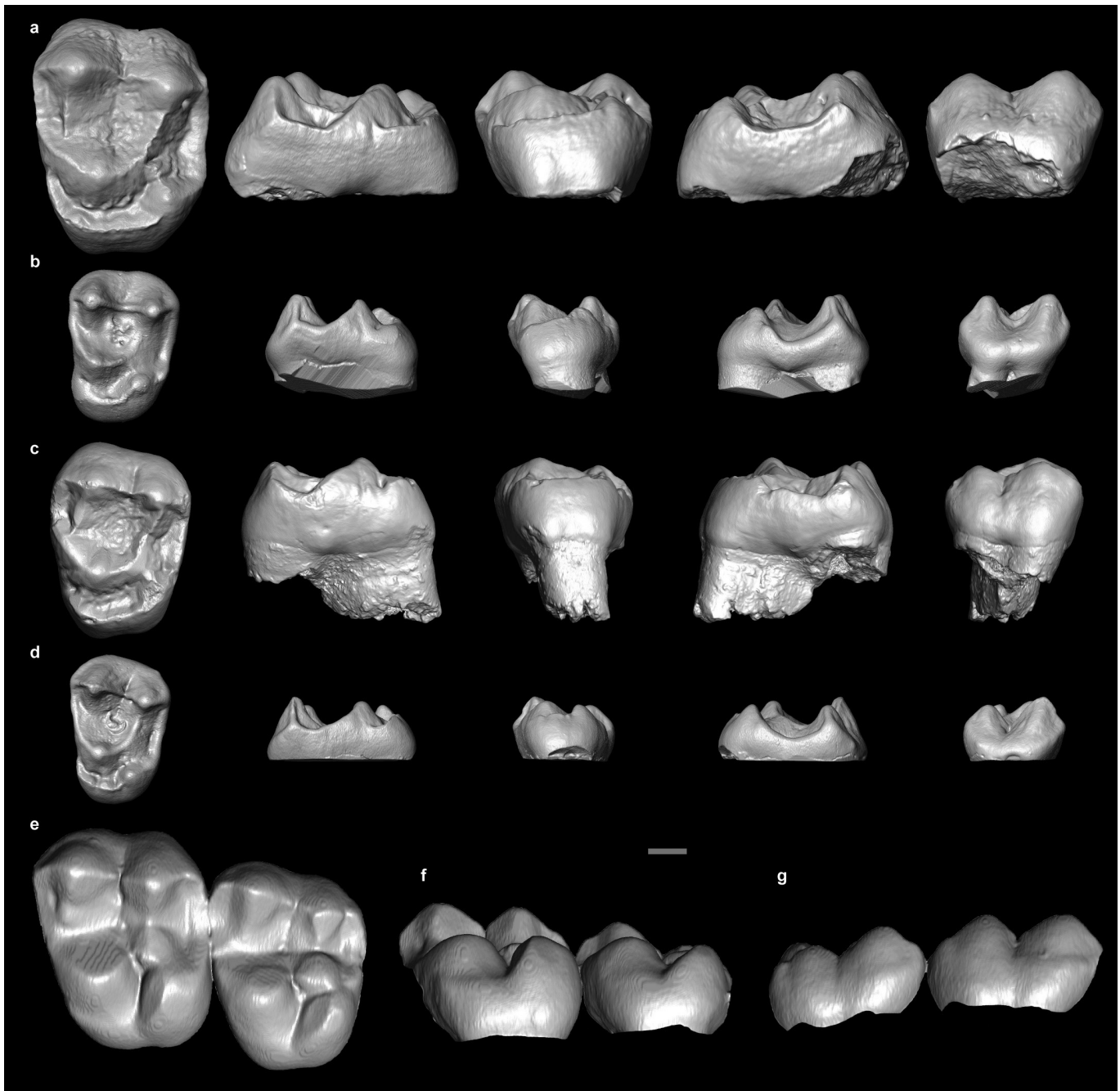
**Extended Data Figure 3 | Generic-level biostratigraphy of selected taxa from the early Miocene (~21 Ma) Las Cascadas Formation, Panama.** Individual biochronologies were interpreted from temporal ranges known from higher-latitudes. The calibration of the NALMA boundaries is as described previously<sup>13,14</sup>.



**Extended Data Figure 4 | Detailed upper molar comparisons of *Panamacebus* to *Neosaimiri*.** a–d, Left to right, occlusal, distal, lingual and buccal views of the left M<sup>1</sup> (UF 280128: <http://dx.doi.org/10.17602/M2/M8531>) (a) and the left M<sup>2</sup> (UF 281001: <http://dx.doi.org/10.17602/M2/M8550>) (d) of *P. transitus* compared with M<sup>1</sup> and M<sup>2</sup> of *N. fieldsi* (b, c, e–g): right M<sup>1</sup> (IGM-KU 89008: <http://dx.doi.org/10.17602/M2/M8541>) (b), right M<sup>1</sup> (IGM 89019:

<http://dx.doi.org/10.17602/M2/M8544>) (c), right M<sup>2</sup> (IGM-KU 89018: <http://dx.doi.org/10.17602/M2/M8543>) (e), right M<sup>2</sup> (IGM-KU 89104: <http://dx.doi.org/10.17602/M2/M8548>) (f), and right M<sup>2</sup> (IGM-KU 89011: <http://dx.doi.org/10.17602/M2/M8542>) (g). Images were generated from microCT scan data (see Methods and links associated with specimen numbers). Right-sided teeth were flipped to facilitate comparison with left-sided teeth. Scale bar, 1 mm.

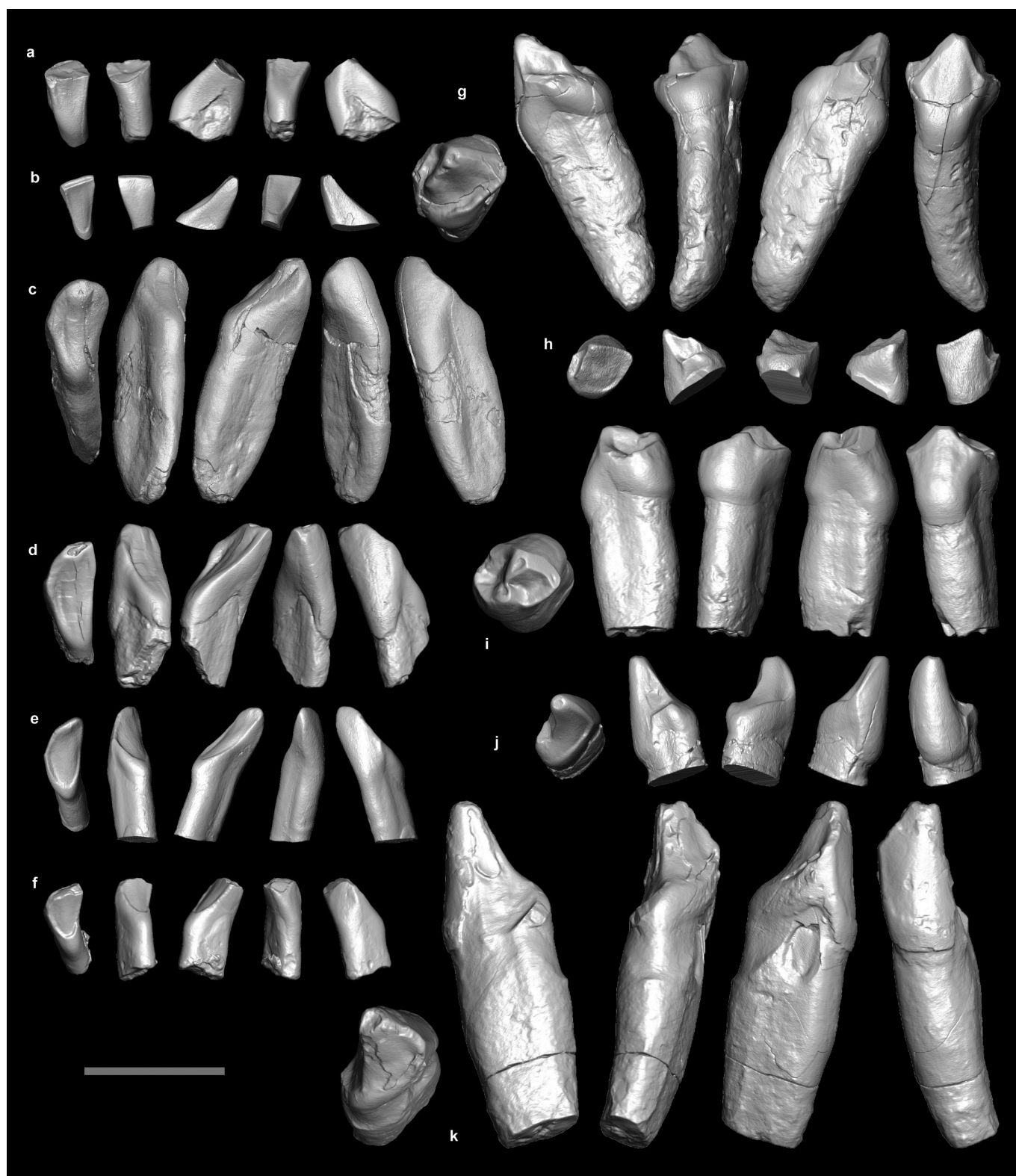




**Extended Data Figure 5 | Detailed upper molar comparisons of *Panamacebus* to *Neosaimiri* and *Cebus*.** a–g, Left to right, occlusal, distal, lingual, mesial and buccal views of the left M<sup>1</sup> (UF 280128; <http://dx.doi.org/10.17602/M2/M8531>) (a) and the left M<sup>2</sup> (UF 281001; <http://dx.doi.org/10.17602/M2/M8550>) (c) of *P. transitus* compared with M<sup>1</sup> and M<sup>2</sup> of *N. fieldsi* (b, d); right M<sup>1</sup> (IGM 89019; <http://dx.doi.org/10.17602/M2/M8544>) (b), and right M<sup>2</sup> (IGM-KU 89104; <http://dx.doi.org/10.17602/>

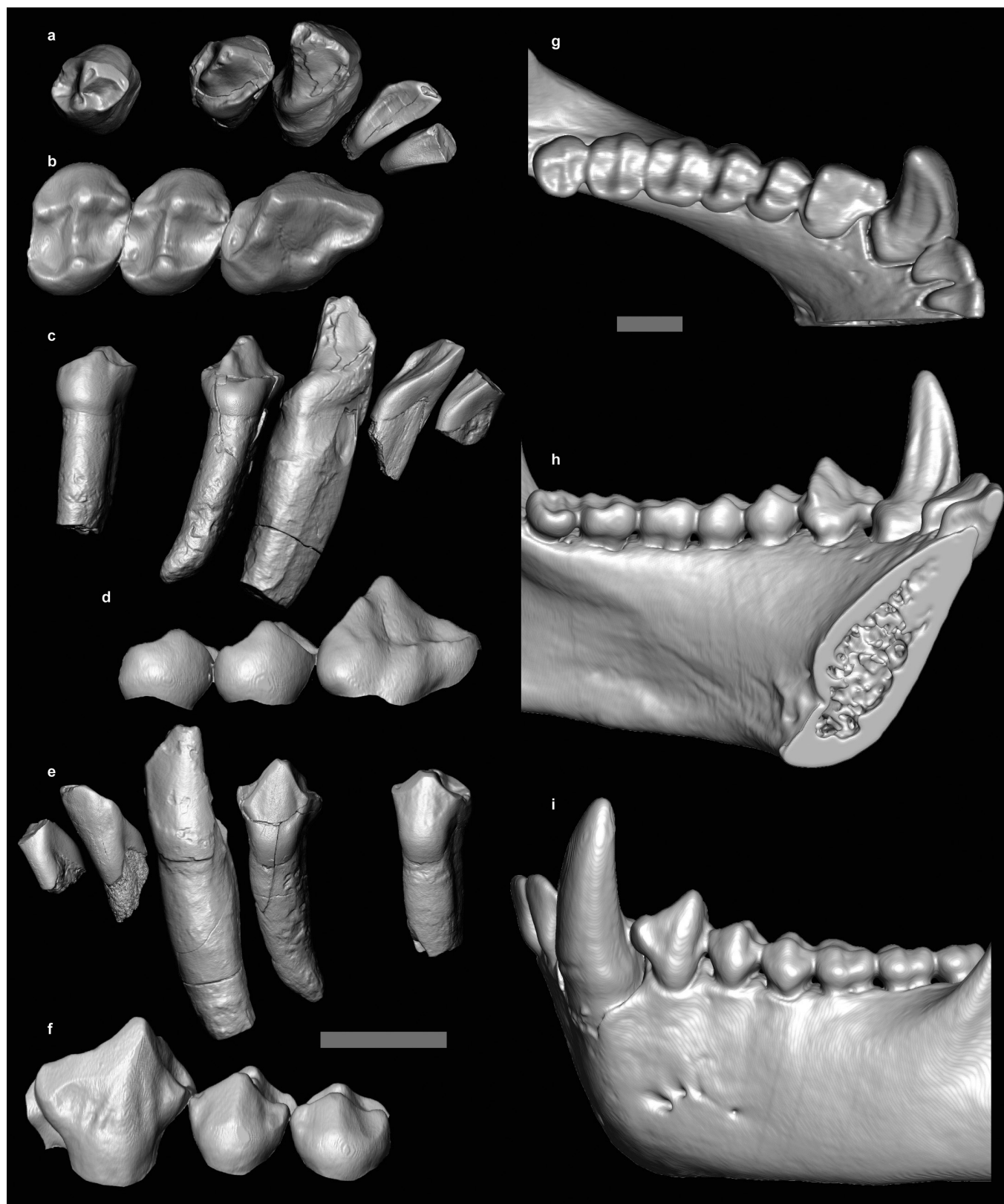
M2/M8548) (d); and the left M<sup>1</sup> and M<sup>2</sup> of *Cebus capucinus* (USNM 291128; <http://dx.doi.org/10.17602/M2/M8627>) in occlusal (e), lingual (f) and buccal views (g). Images were generated from microCT scan data (see Methods and links associated with specimen numbers). Right-sided teeth were flipped to facilitate comparison with left-sided teeth. Scale bar, 1 mm.





**Extended Data Figure 6 | Detailed comparisons of *Panamacebus* lower teeth to those of *Neosaimiri* and *Stirtonia*.** a–k, Left to right, occlusal, distal, lingual, mesial and buccal views of the partial left I<sub>1</sub> (UF 280130: <http://dx.doi.org/10.17602/M2/M8400>) (a), right I<sub>2</sub> (UF 267048, mirrored: <http://dx.doi.org/10.17602/M2/M8395>) (d), left P<sub>2</sub> (UF 280127: <http://dx.doi.org/10.17602/M2/M8397>) (g), left P<sub>4</sub> (UF 280129: <http://dx.doi.org/10.17602/M2/M8399>) (i) and right C<sub>1</sub> (UF 280131, mirrored: <http://dx.doi.org/10.17602/M2/M8401>) (k) of *P. transitus* compared with the right I<sub>1</sub> (IGM-KU 89086: <http://dx.doi.org/10.17602/M2/M8546>) (b), right I<sub>2</sub> (IGM-KU 89092:

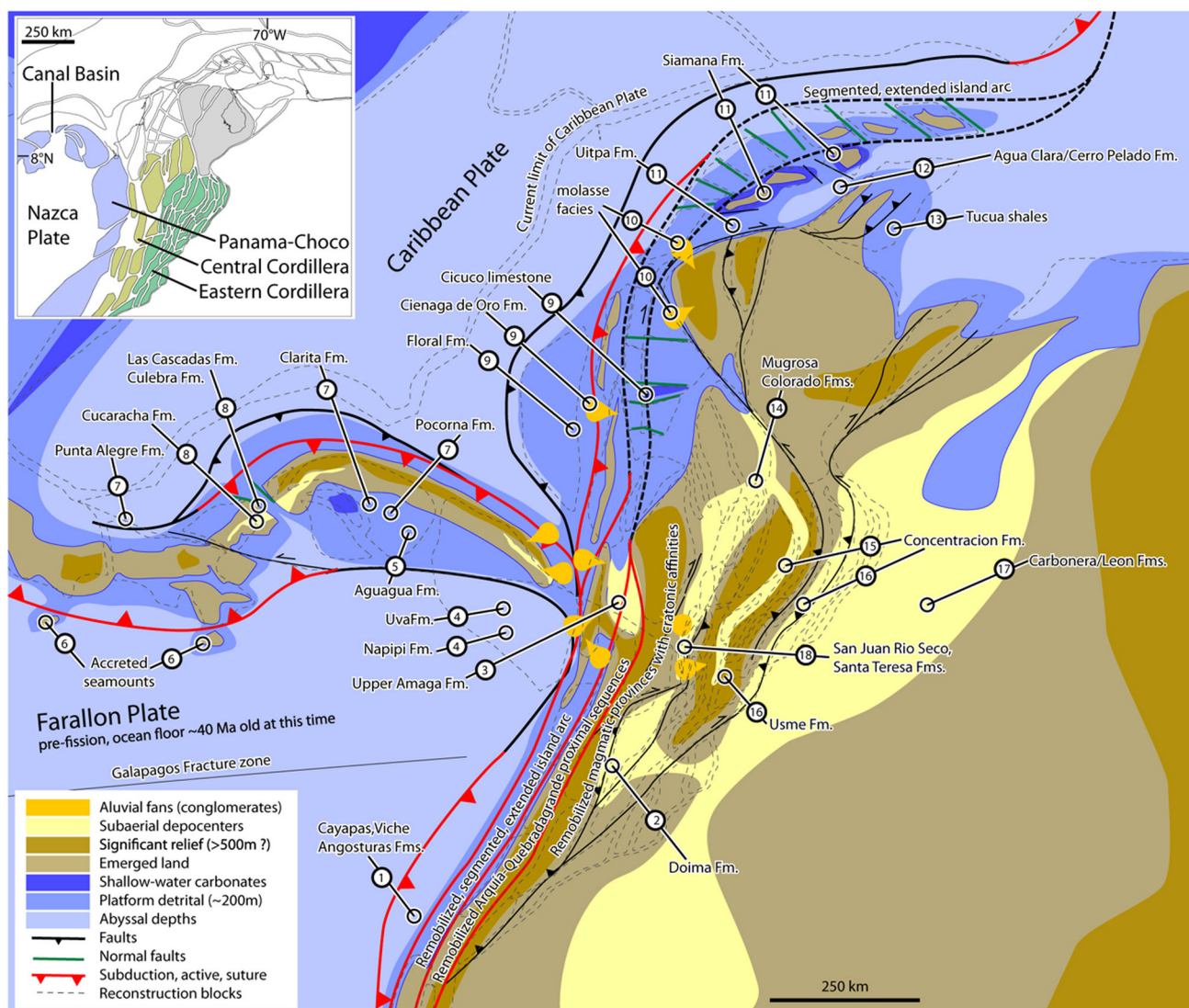
<http://dx.doi.org/10.17602/M2/M8547>) (e), left I<sub>2</sub> (UCMP 39205: <http://dx.doi.org/10.17602/M2/M1880>) (f), right P<sub>2</sub> (IGM-KU 90016: <http://dx.doi.org/10.17602/M2/M8549>) (h) and right C<sub>1</sub> (IGM-KU 89021: <http://dx.doi.org/10.17602/M2/M8432>) (j) of *N. fieldsi* and the right I<sub>2</sub> (UCMP 38989: <http://dx.doi.org/10.17602/M2/M1799>) of *Stirtonia tatarcoensis* (c). Images were generated from microCT scan data (see Methods and links associated with specimen numbers). Right-sided teeth were flipped to facilitate comparison with left-sided teeth. Scale bar, 5 mm.



**Extended Data Figure 7 | Detailed comparisons of *Panamacebus* lower teeth to those of *Cebus*.** a–i, Occlusal (a), lingual (c) and buccal (e) views of *P. transitus* partial left  $I_1$  (UF 280130: <http://dx.doi.org/10.17602/M2/M8400>), right  $I_2$  (UF 267048, mirrored: <http://dx.doi.org/10.17602/M2/M8395>), right  $C_1$  (UF 280131, mirrored: <http://dx.doi.org/10.17602/M2/M8401>), left  $P_2$  (UF 280127: <http://dx.doi.org/10.17602/M2/M8397>) and left  $P_4$  (UF 280129: <http://dx.doi.org/10.17602/M2/M8399>); occlusal (b),

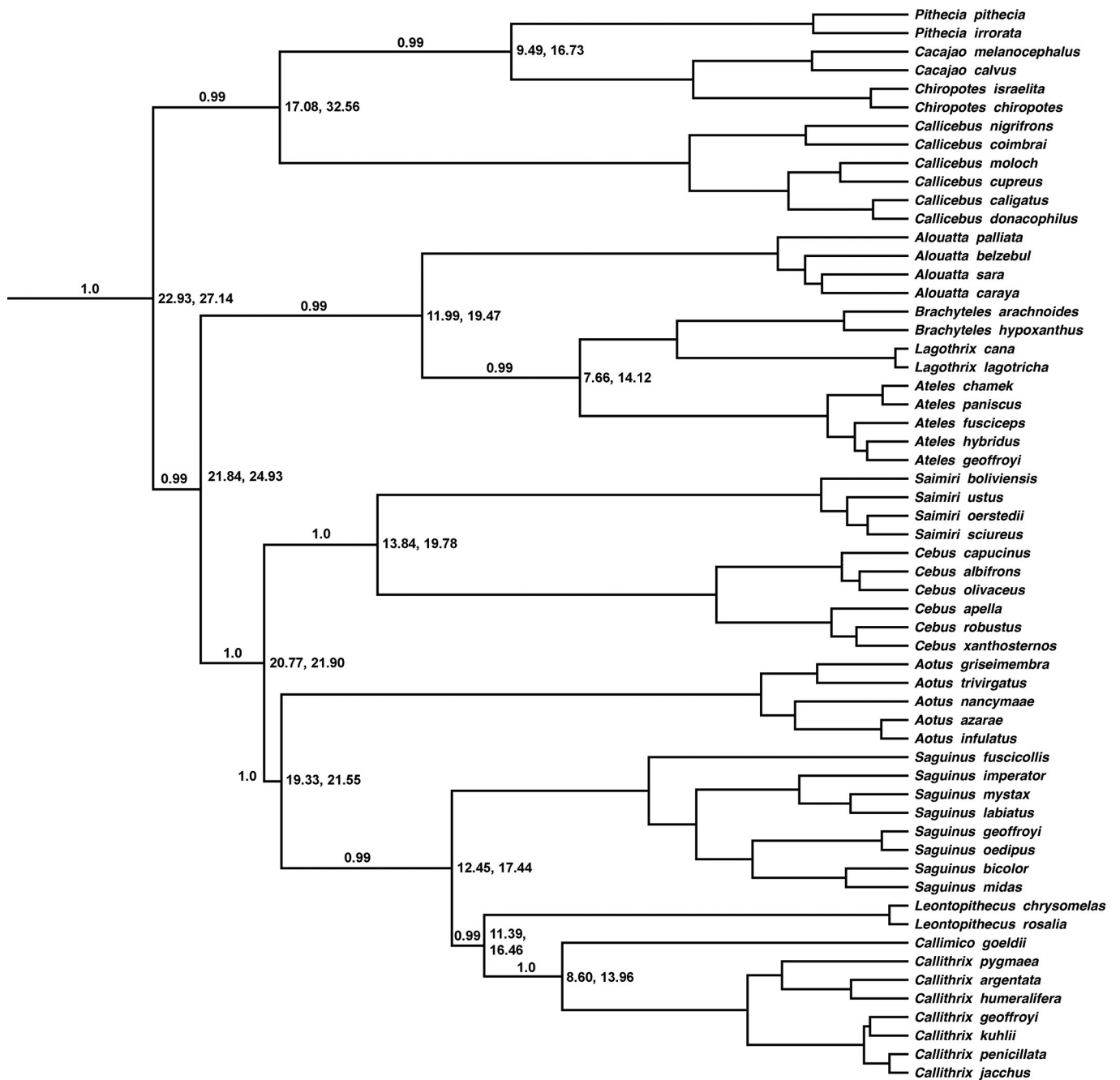
lingual (d) and buccal (f) views of *C. capucinus* right  $P_2$ – $P_4$  (USNM 291128: <http://dx.doi.org/10.17602/M2/M8629>); and occlusal (g), lingual (h) and buccal (i) views of *C. capucinus* left dentary with  $I_1$ – $M_3$  (USNM 291236: <http://dx.doi.org/10.17602/M2/M8622>). Right-sided teeth were mirrored to facilitate comparison with left-sided teeth. Images were generated from microCT scan data (see Methods). Scale bars, 5 mm.





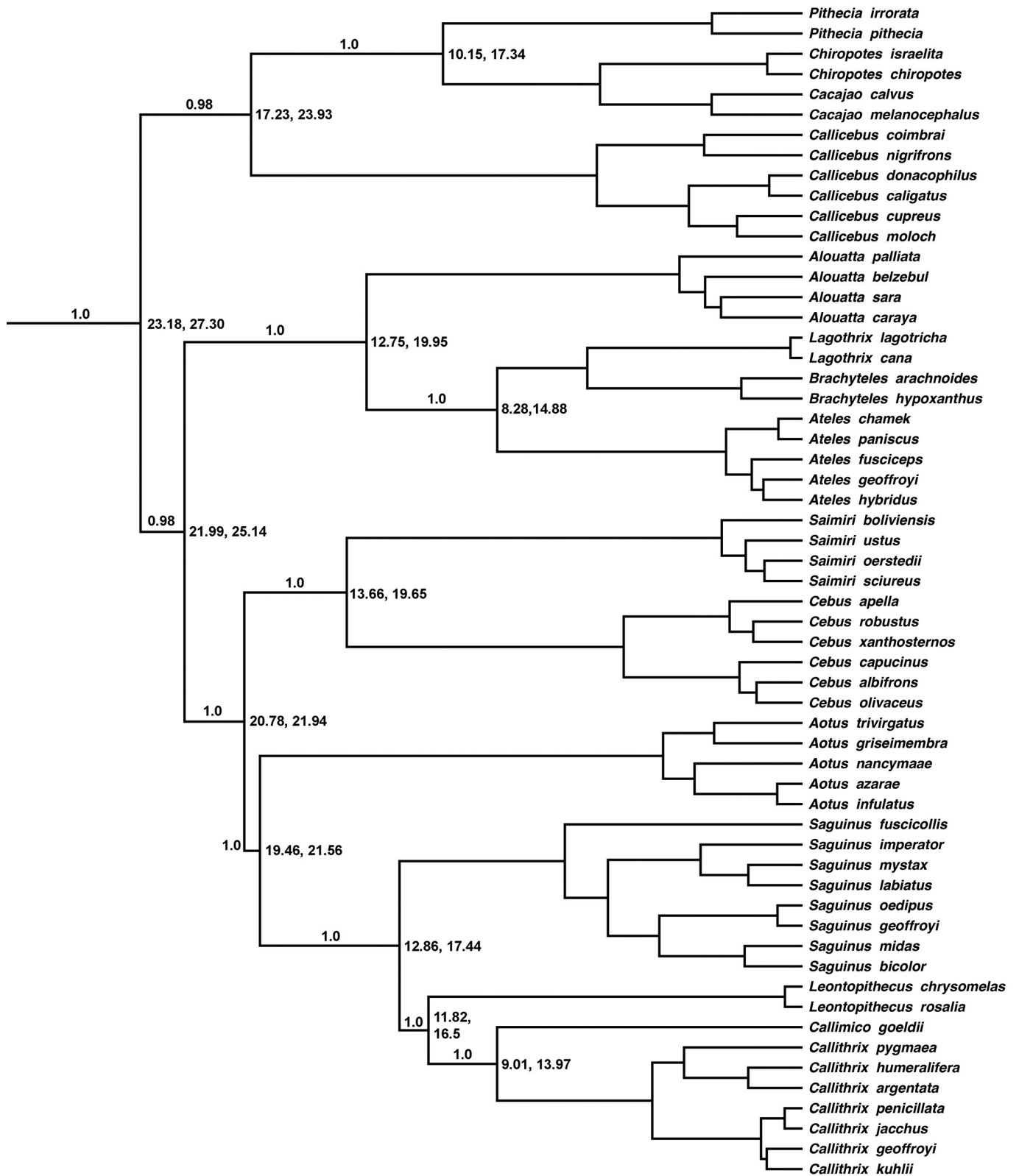
**Extended Data Figure 8 | Detailed palaeogeographic reconstruction of the Panama Canal Basin region during the late Oligocene–early Miocene, showing the location of key geological formations, faults and tectonic blocks.** Sedimentary environments were extrapolated from published stratigraphic sections that were placed over the palinspastic reconstruction in the following locations: (1) Pacific sections that include conglomerate and sandy strata; (2) upper Magdalena Basin; (3) Amaga Formation coal-bearing and sandy-conglomeratic; (4) Choco;

(5) easternmost Panama; (96) western Panama; (7) Panama; (8) Canal basin; (9) northwestern Colombia; (10) Sierra Nevada Santa Marta, unnamed sandy and conglomeratic strata; (11) Guajira Peninsula; (12) Falcon; (13) Falcon/Lara; (14) middle Magdalena Basin; (15) Floresta Massif; (16) axial Cordillera Oriental; (17) foothills; (18) southern middle Magdalena Basin. See Supplementary Methods for references and detailed discussion.



**Extended Data Figure 9 |** Maximum clade credibility tree summarizing the concatenated trees from the divergence dating analysis using the birth-death model showing the Platyrrhini. Ninety-five per cent highest posterior density intervals are shown on key nodes and the corresponding branches are labelled with posterior probability values. For clarity, only the Platyrrhini are shown here.





Extended Data Figure 10 | Maximum clade credibility tree summarizing the concatenated trees from divergence dating the analysis using the Yule model showing the Platyrrhini. Ninety-five per cent highest posterior density intervals are shown on key nodes and the corresponding branches are labelled with posterior probability values. For clarity, only the Platyrrhini are shown here.

# Restoring cortical control of functional movement in a human with quadriplegia

Chad E. Bouton<sup>1†</sup>, Ammar Shaikhouni<sup>2,3</sup>, Nicholas V. Annetta<sup>1</sup>, Marcia A. Bockbrader<sup>2,4</sup>, David A. Friedenberg<sup>5</sup>, Dylan M. Nielson<sup>2,3</sup>, Gaurav Sharma<sup>1</sup>, Per B. Sederberg<sup>2,6</sup>, Bradley C. Glenn<sup>7</sup>, W. Jerry Mysiw<sup>2,4</sup>, Austin G. Morgan<sup>1</sup>, Milind Deogaonkar<sup>2,3</sup> & Ali R. Rezai<sup>2,3</sup>

Millions of people worldwide suffer from diseases that lead to paralysis through disruption of signal pathways between the brain and the muscles. Neuroprosthetic devices are designed to restore lost function and could be used to form an electronic ‘neural bypass’ to circumvent disconnected pathways in the nervous system. It has previously been shown that intracortically recorded signals can be decoded to extract information related to motion, allowing non-human primates and paralysed humans to control computers and robotic arms through imagined movements<sup>1–11</sup>. In non-human primates, these types of signal have also been used to drive activation of chemically paralysed arm muscles<sup>12,13</sup>. Here we show that intracortically recorded signals can be linked in real-time to muscle activation to restore movement in a paralysed human. We used a chronically implanted intracortical microelectrode array to record multiunit activity from the motor cortex in a study participant with quadriplegia from cervical spinal cord injury. We applied machine-learning algorithms to decode the neuronal activity and control activation of the participant’s forearm muscles through a custom-built high-resolution neuromuscular electrical stimulation system. The system provided isolated finger movements and the participant achieved continuous cortical control of six different wrist and hand motions. Furthermore, he was able to use the system to complete functional tasks relevant to daily living. Clinical assessment showed that, when using the system, his motor impairment improved from the fifth to the sixth cervical (C5–C6) to the seventh cervical to first thoracic (C7–T1) level unilaterally, conferring on him the critical abilities to grasp, manipulate, and release objects. This is the first demonstration to our knowledge of successful control of muscle activation using intracortically recorded signals in a paralysed human. These results have significant implications in advancing neuroprosthetic technology for people worldwide living with the effects of paralysis.

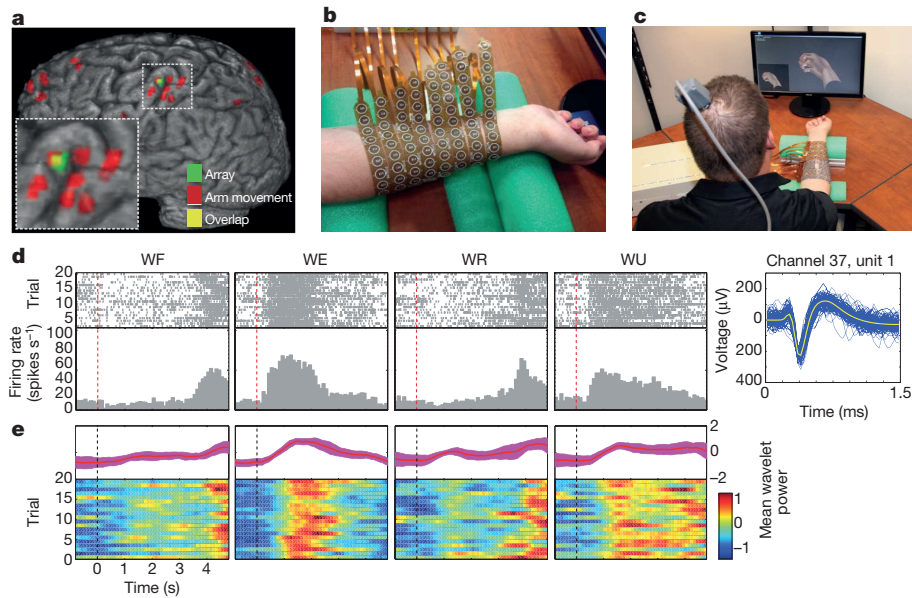
The study participant was a 24-year-old male with stable, non-spastic C5/C6 quadriplegia from cervical spinal cord injury (SCI) sustained in a diving accident 4 years previously. He underwent implantation of a Utah microelectrode array (Blackrock Microsystems) in his left primary motor cortex. As shown in Fig. 1a, the hand area of the primary motor cortex was identified preoperatively by performing functional magnetic resonance imaging (fMRI) while the participant attempted to mirror videos of hand movements. The final array implantation location was chosen during surgery, targeting the hand area while avoiding sulci and injury to large cortical vessels. The implant location was confirmed by co-registration of postoperative computed tomography imaging with preoperative fMRI (Fig. 1a) and is consistent with the ‘knob’ region of the primary motor cortex<sup>5,14</sup>.

The participant attended up to three sessions weekly for 15 months after implantation to use the neural bypass system (NBS). In each session, he was trained to utilize his motor cortical neuronal activity to control a custom-built high-resolution neuromuscular electrical stimulator (NMES). The NMES delivered electrical stimulation to his paralysed right forearm muscles using an array of 130 electrodes embedded in a custom-made flexible sleeve wrapped around the arm (Fig. 1b). The participant was positioned in front of a computer monitor, and a stereo camera was placed overhead to track and record hand movements (Fig. 1c). During the study, up to 50 single units could be isolated in a given session. Near the end of the study, 33 units could be isolated with a mean signal-to-noise ratio of  $3.05 \pm 0.81$  (mean  $\pm$  s.d.) including units that responded to imagined or performed wrist movements (Fig. 1d). (See Extended Data Fig. 1 for additional unit activity.) Wavelet decomposition of the multiunit activity recorded from 96 microelectrodes was used to produce mean wavelet power (MWP) features for decoding (Fig. 1e) (see Methods).

To assess the ability of the NBS to restore individual movements, we focused on six wrist and hand movements that were all impaired by the participant’s injury and reactivated by stimulation of forearm muscles (see Supplementary Video 1 showing the participant attempting the six movements without the use of the NBS). Each session began with recalibration of the NMES to map electrode stimulation patterns to evoked movements (see Methods). Cortical activity was continuously decoded as the participant attempted the six selected movements interleaved with rest periods, as cued by an animated virtual hand on the computer monitor. Changes in the MWP patterns for each movement were captured during the test. These patterns were then processed by multiple simultaneous neural decoders, one for each trained movement, using a nonlinear kernel method with a non-smooth support vector machine<sup>15</sup>. The decoders were trained in successive blocks and, once trained, their outputs were continuously compared using the highest decoder output to control the corresponding NMES movement stimulation pattern (see Methods). During movement, a large portion of the stimulation artefact that occurred during a stimulation pulse was removed, but stimulation effects still remained (see Methods).

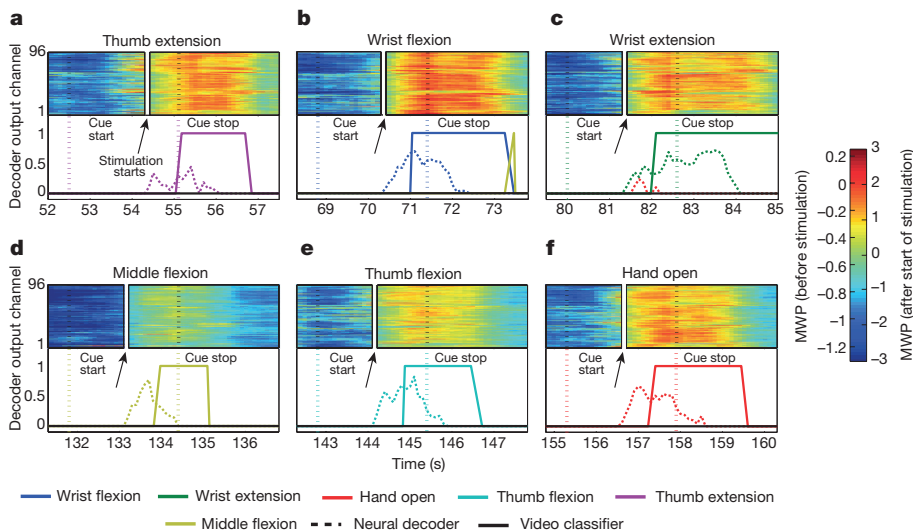
To test the system’s performance, test blocks were performed consisting of five trials of each of the six trained movements presented in random order. At the beginning of each trial, the participant was visually cued by the virtual hand demonstrating a target movement. Representative data, including modulation of MWP (before and after stimulation begins), decoder outputs, and corresponding movement state are shown in Fig. 2. MWP increases by a factor of 2–8 after stimulation begins because of residual stimulation artefact (see Methods and Extended Data Fig. 2). However, since the neural decoders were trained with MWP from before and during stimulation, they were able

<sup>1</sup>Medical Devices and Neuromodulation, Battelle Memorial Institute, 505 King Avenue, Columbus, Ohio 43201, USA. <sup>2</sup>Center for Neuromodulation, The Ohio State University, Columbus, Ohio 43210, USA. <sup>3</sup>Department of Neurological Surgery, The Ohio State University, Columbus, Ohio 43210, USA. <sup>4</sup>Department of Physical Medicine and Rehabilitation, The Ohio State University, Columbus, Ohio 43210, USA. <sup>5</sup>Advanced Analytics and Health Research, Battelle Memorial Institute, 505 King Avenue, Columbus, Ohio 43201, USA. <sup>6</sup>Department of Psychology, The Ohio State University, Columbus, Ohio 43210, USA. <sup>7</sup>Energy Systems, Battelle Memorial Institute, 505 King Avenue, Columbus, Ohio 43201, USA. <sup>†</sup>Present address: Feinstein Institute for Medical Research, 350 Community Drive, Manhasset, New York 11030, USA.



**Figure 1 | Experimental setup and neural modulation.** **a**, Red regions are brain areas active during attempts to mimic hand movements, where the  $t$ -values for the move-rest T1-weighted fMRI contrast are greater than 7; The implanted microelectrode array location from post-operation computed tomography is shown in green; The overlap of the red and green regions is shown in yellow. **b**, Neuromuscular electrical stimulation sleeve. **c**, NBS in use with the participant in front a computer monitor. **d**, Examples of rasters and peristimulus histograms from one discriminated unit (channel 37 unit 1) with response to attempted wrist movements. The participant was presented with cues to attempt wrist flexion (WF), wrist extension (WE), wrist radial deviation (WR), and wrist ulnar deviation (WU). Each cue was presented for a random duration of 3–4 s followed by a random 3–4 s rest period. We presented 20 trials of each in random order. The top part of each subpanel is a raster, each dot represents a spike, and each row of spikes represents data from one trial. All trials were aligned on cue presentation (time zero, red dashed line). At the right of

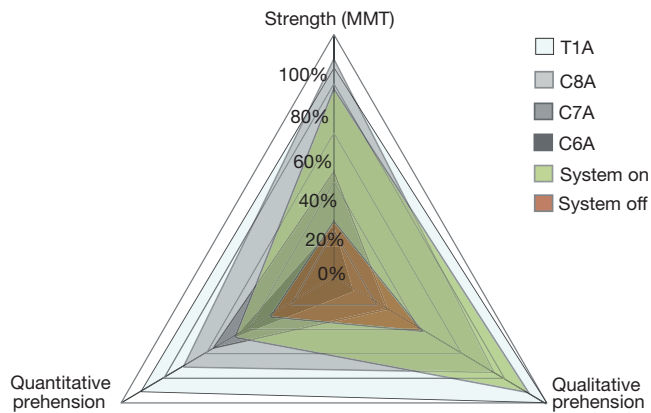
the panel is an example of 500 randomly selected waveforms from the discriminated unit (blue) from channel 37 that responded to wrist, elbow, and shoulder movements. The thick yellow line represents the average waveform for this unit which showed increased firing rate for imagined wrist flexion and wrist ulnar deviation (as well as elbow and shoulder movements, see Extended Data Fig. 1). Because the participant was asked to imagine and hold the movement throughout the cue period, the increase of activity during the rest cue is probably due to the participant imagining the antagonist movement to return to a neutral position. **e**, Wavelet-processed neural data from channel 37 corresponding to the same movements in **d**. The bottom part of each subpanel shows the mean wavelet coefficients for each movement and each trial over time. Each row represent data from a trial. The top part shows the MWP averaged over the trials in dark red with a confidence interval of 1 s.d. around the mean shown in light pink. The trials were centred on cue presentation (black dashed line). (Photographs by A. Morgan and T. R. Massey.)



**Figure 2 | MWP and system performance for individual hand movements.** **a–f**, Heat maps of MWP representing processed signals from the microelectrode array, neural decoder output scores (dashed line), and physical hand movements (solid line) as detected by a computer-based video classification algorithm for each movement type: thumb extension (**a**), wrist flexion (**b**), wrist extension (**c**), middle flexion (**d**), thumb flexion (**e**), and hand open (**f**). Of the six simultaneously running decoders, the output score from the one with the highest amplitude greater than zero was used to turn on/off the stimulation. The vertical white line marks when the decoder crosses zero and the stimulation is turned on. MWP increases (by a factor of 2–8) during stimulation because of residual

stimulation artefact (see Methods). The MWP plot is split into two areas with two different colour scales to facilitate visualization of MWP before and during stimulation. Trial data shown range from 0.5 s before the cue to 2.5 s after the cue ended. A 1 s wide boxcar filter is applied to the MWP data causing a delay of  $\sim 0.5$  s. Each subfigure represents one trial out of the total five trials for each individual hand movement and the time scale on the  $x$  axis shows the absolute time when the particular movement occurred during the test block. The MWP is in units of standard deviations away from a baseline non-movement period. Only decoder outputs greater than zero are shown for visual clarity.





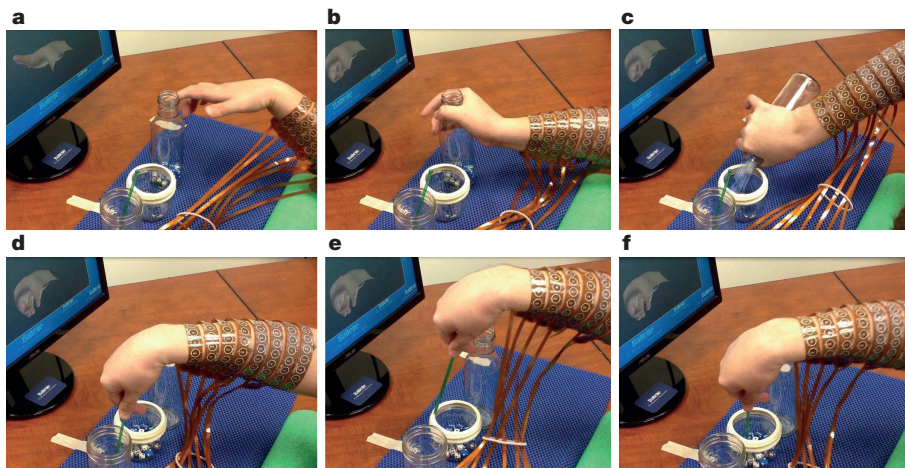
**Figure 3 | GRASSP clinical assessment.** GRASSP performance on each domain was scored by percentage of possible points earned, and normed to benchmarks of the International Standards for Neurological Classification of Spinal Cord Injury and the American Spinal Injury Association Impairment Scale (greyscale regions in graph). The participant's baseline score (red), obtained 8 months after implantation, falls within the C5–C6 norms for strength on MMT and quantitative prehension (gross grasping ability) and between C7–C8 for qualitative prehension (prehensile skills). Using the NBS (green), between 8 and 9 months after implantation, his motor function improved in all areas tested, with strength nearly achieving C8–T1 level function, qualitative prehension similarly improving to between C8–T1 level, and quantitative prehension improving to C6 level. Qualitative prehension was limited by lack of stimulation electrodes to facilitate thumb opposition. MMT strength was limited by lack of stimulation electrodes over elbow extensors and hand intrinsic.

to recognize the correct imagined movement to initiate stimulation and the participant's desire to sustain and subsequently terminate the target movement (see Supplementary Video 2 showing the participant attempting the six movements using NBS). Accuracy (percentage of video frames where observed movement matches the cue), sensitivity (percentage of video frames during cued movement where correct movement was observed), and specificity (percentage of video frames correctly identified during rest or non-target movement cues) for the trained movements were measured by computer-based evaluation of video frames of hand movements (see Methods and Extended Data Fig. 3a). Using the system, the participant achieved an overall accuracy of  $70.4 \pm 1.0\%$  (mean  $\pm$  s.d.,  $P < 0.01$  by permutation test<sup>16</sup>). His accuracy for performing individual movements ranged from

$93.1 \pm 0.5\%$  for wrist flexion ( $P < 0.01$ ) to  $97.3 \pm 0.3\%$  for thumb flexion ( $P < 0.01$ ). Sensitivity ranged between  $32.9 \pm 3.8\%$  for thumb extension and  $81.9 \pm 2.6\%$  for wrist extension, and specificity ranged between  $94.8 \pm 0.5\%$  for wrist flexion and  $99.8 \pm 1.0\%$  for thumb flexion (Extended Data Fig. 3b). The participant was able to perform similar basic movement tasks with text or verbal cues as well, suggesting the mirror neuron system is not required for these tasks.

To characterize the recovery of upper limb function when using the NBS, a physiatrist performed assessments of upper limb sensorimotor function using the Graded and Redefined Assessment of Strength, Sensibility, and Prehension (GRASSP) test<sup>17</sup>. The GRASSP is a standardized test with excellent inter-rater and test–retest reliability developed to assess sensorimotor impairment of patients with cervical SCI across the following domains of the Manual Muscle Test (MMT): strength, dorsal sensation, ventral sensation, gross grasping ability (qualitative prehension), and prehensile skills (quantitative prehension). All five GRASSP domains were assessed without use of NBS to generate a baseline score. As only motor, and not sensory, function was expected to change with use of the NBS, MMT, gross grasping ability, and prehensile skills subtests were repeated using the system. The NMES was calibrated to map electrode stimulation patterns to desired evoked movement patterns using an anatomically guided trial-and-error approach and a new decoder was trained for each individual task, rather than one for all tasks (see Methods). Figure 3 shows that when the participant used the NBS, his MMT strength improved from C6 to C7–C8 level, his gross grasping ability improved from C7–C8 to C8–T1 level, and his prehensile skills improved from C5 to C6 level. These findings quantify the reduction in functional motor impairments possible with NBS for patients with cervical SCI. Fatigue was sometimes observed after multiple repetitions of sustained (3–4 s) stimulation.

To investigate the effectiveness of the NBS in assisting with daily activities relevant to daily life, we asked the participant to perform a complex functional movement that he was unable to complete without the system. The task required him to grasp a bottle with a cylindrical grasp, maintain the grasp while pouring its contents into a jar, open his hand to release the bottle, and use a precision pinch to pick up and hold a stir stick to stir the jar's contents (Fig. 4 and Supplementary Video 3). During the task, the participant used his residual shoulder and elbow movements to guide his hand in space while he performed the task; however, this presented a significant challenge in initial testing. The participant had difficulty in maintaining his grasp of the objects, particularly during transfer, and often dropped them. We observed differences in the MWP associated with imagined grasping,



**Figure 4 | Grasp-pour-and-stir functional movement task.** a–f, Sequential snapshots from the functional movement task showing the participant opening his hand (a), grasping the glass bottle (b), pouring its contents (dice) into a jar (c), grasping a stir stick from another jar (d), transferring the stir stick without dropping it (e), and using it to stir the

dice in the jar (f). The task required the participant to evoke different cortical modulation patterns to control the opening of his hand, perform a cylindrical palmar grasp, and achieve a precision pinch grasp, while simultaneously moving his arm. (Video by C. Majstorovic.)



stimulation-induced grasping, and grasp-and-transfer movements (Extended Data Fig. 4). To improve the decoding performance during these complex tasks, the MWP data from all three of these periods were used to train a nonlinear support vector machine (see Methods). Using this approach, the participant was able to successfully complete the grasp-pour-and-stir task three out of five times in 10 min with a completion time of  $42 \pm 10$  s (mean  $\pm$  s.d.) for successful trials, demonstrating that he was able to utilize the NBS for a functional activity that he could not perform otherwise (see Supplementary Video 4 showing the participant attempting the task without the use of NBS).

In this study, for the first time, a human with quadriplegia regained volitional, functional movement through the use of intracortically recorded signals linked to neuromuscular stimulation in real-time. With use of the investigational system, our C5/C6 participant gained wrist and hand function consistent with a C7–T1 level of injury. This improvement in function is meaningful for reducing the burden of care in patients with SCI as most C5 and C6 patients require assistance for activities of daily living, while C7–T1 level patients can live more independently. Although invasive, the NBS provides an advantage over existing functional electrical stimulation systems that utilize low dimensional control signals such as EEG or EMG like the Freehand system<sup>18–20</sup>. These devices typically allow control over fewer movements than those demonstrated in this study, because of the relatively low information content of their control signal sources compared with intracortically recorded signals.

Complex functional movements that involve both stimulation-induced and residual movements, such as those at the shoulder and elbow, presented an important challenge in this study. Residual limb movement can activate motor cortical neurons<sup>21</sup> and some motor cortical neurons respond to movement from several joints<sup>22</sup>. This may explain the neural activity differences observed and the difficulties initially encountered as the patient attempted movements involving paralysed and non-paralysed joint movement simultaneously. The nonlinear algorithm presented here is one possible solution to this problem. To allow transfer of this work to other patients, further improvements will be required on the microelectrode technology, algorithms, and NMES. However, the electronic neural bypass presented here demonstrates what is possible in the future and can offer hope for movement restoration to people living with paralysis worldwide.

**Online Content** Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

**Received 19 August 2015; accepted 15 February 2016.**

**Published online 13 April 2016.**

1. Aflalo, T. *et al.* Decoding motor imagery from the posterior parietal cortex of a tetraplegic human. *Science* **348**, 906–910 (2015).
2. Bansal, A. K., Truccolo, W., Vargas-Irwin, C. E. & Donoghue, J. P. Decoding 3D reach and grasp from hybrid signals in motor and premotor cortices: spikes, multiunit activity, and local field potentials. *J. Neurophysiol.* **107**, 1337–1355 (2012).
3. Chapin, J. K., Moxon, K. A., Markowitz, R. S. & Nicolelis, M. A. Real-time control of a robot arm using simultaneously recorded neurons in the motor cortex. *Nature Neurosci.* **2**, 664–670 (1999).
4. Hochberg, L. R. *et al.* Reach and grasp by people with tetraplegia using a neurally controlled robotic arm. *Nature* **485**, 372–375 (2012).

5. Hochberg, L. R. *et al.* Neuronal ensemble control of prosthetic devices by a human with tetraplegia. *Nature* **442**, 164–171 (2006).
6. Kennedy, P. R. & Bakay, R. A. Restoration of neural output from a paralyzed patient by a direct brain connection. *Neuroreport* **9**, 1707–1711 (1998).
7. Santhanam, G., Ryu, S. I., Yu, B. M., Afshar, A. & Shenoy, K. V. A high-performance brain-computer interface. *Nature* **442**, 195–198 (2006).
8. Serruya, M. D., Hatsopoulos, N. G., Paninski, L., Fellows, M. R. & Donoghue, J. P. Instant neural control of a movement signal. *Nature* **416**, 141–142 (2002).
9. Taylor, D. M., Tillery, S. I. & Schwartz, A. B. Direct cortical control of 3D neuroprosthetic devices. *Science* **296**, 1829–1832 (2002).
10. Velliste, M., Perel, S., Spalding, M. C., Whitford, A. S. & Schwartz, A. B. Cortical control of a prosthetic arm for self-feeding. *Nature* **453**, 1098–1101 (2008).
11. Wessberg, J. *et al.* Real-time prediction of hand trajectory by ensembles of cortical neurons in primates. *Nature* **408**, 361–365 (2000).
12. Ethier, C., Oby, E. R., Bauman, M. J. & Miller, L. E. Restoration of grasp following paralysis through brain-controlled stimulation of muscles. *Nature* **485**, 368–371 (2012).
13. Moritz, C. T., Perlmutter, S. I. & Fetz, E. E. Direct control of paralysed muscles by cortical neurons. *Nature* **456**, 639–642 (2008).
14. Yousry, T. A. *et al.* Localization of the motor hand area to a knob on the precentral gyrus. A new landmark. *Brain* **120**, 141–157 (1997).
15. Humber, C., Ito, K. & Bouton, C. Nonsmooth formulation of the support vector machine for a neural decoding problem. *arXiv Preprint* at <http://arxiv.org/abs/1012.0958v1> (2010).
16. Ojala, M. & Garriga, G. Permutation tests for studying classifier performance. *J. Mach. Learn. Res.* **11**, 1833–1863 (2010).
17. Kalsi-Ryan, S., Curt, A., Verrier, M. C. & Fehlings, M. G. Development of the Graded Redefined Assessment of Strength, Sensibility and Prehension (GRASP): reviewing measurement specific to the upper limb in tetraplegia. *J. Neurosurg. Spine* **17**, 65–76 (2012).
18. Kilgore, K. L. *et al.* An implanted upper-extremity neuroprosthesis using myoelectric control. *J. Hand Surg. Am.* **33**, 539–550 (2008).
19. Peckham, P. H. *et al.* An advanced neuroprosthesis for restoration of hand and upper arm control using an implantable controller. *J. Hand Surg. Am.* **27**, 265–276 (2002).
20. Pfurtscheller, G., Muller, G. R., Pfurtscheller, J., Gerner, H. J. & Rupp, R. ‘Thought’-control of functional electrical stimulation to restore hand grasp in a patient with tetraplegia. *Neurosci. Lett.* **351**, 33–36 (2003).
21. Shaikhouni, A., Donoghue, J. P. & Hochberg, L. R. Somatosensory responses in a human motor cortex. *J. Neurophysiol.* **109**, 2192–2204 (2013).
22. Fetz, E. E., Finocchio, D. V., Baker, M. A. & Soso, M. J. Sensory and motor responses of precentral cortex cells during comparable passive and active joint movements. *J. Neurophysiol.* **43**, 1070–1089 (1980).

**Supplementary Information** is available in the online version of the paper.

**Acknowledgements** We thank the study participant for his dedication and his family for their support. We also thank the development team and management at Battelle Memorial Institute for their support, the surgical support team, C. Majstorovic for assistance with data analysis and equipment during sessions, S. Preston for stereo camera coding and troubleshooting, the clinical study support staff, W. Pease for performing the electromyogram and nerve conduction studies, and M. Zhang for assistance with figure preparation. Financial support for this study came from Battelle Memorial Institute and The Ohio State University Center for Neuromodulation.

**Author Contributions** C.E.B., N.V.A., D.A.F., G.S., B.C.G., M.A.B., A.S., A.G.M., D.M.N., P.B.S., W.J.M., and A.R.R. conceived and designed the experiments and fMRI procedure. N.V.A., D.A.F., G.S., B.C.G., M.A.B., and A.G.M. performed the experiments. W.J.M. and M.A.B. were involved in participant recruitment. A.R.R., M.D., and A.S. performed the surgery. All authors contributed to writing the paper.

**Author Information** Reprints and permissions information is available at [www.nature.com/reprints](http://www.nature.com/reprints). The authors declare competing financial interests: details are available in the online version of the paper. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to C.E.B. ([boutonce@gmail.com](mailto:boutonce@gmail.com) or [cbouton@northwell.edu](mailto:cbouton@northwell.edu)) or G.S. ([sharmag@battelle.org](mailto:sharmag@battelle.org)).

## METHODS

No statistical methods were used to predetermine sample size. Cues within an experiment were randomized, but experiments themselves were not randomized. The investigators were not blinded to allocation during experiments and outcome assessment.

Approval for this study was obtained from the US Food and Drug Administration (Investigational Device Exemption) and The Ohio State University Medical Center Institutional Review Board (Columbus, Ohio). The study met institutional requirements for the conduct of human subjects and was registered on the <http://www.ClinicalTrials.gov> website (identifier NCT01997125). The participant referenced in this work provided permission for photographs and videos and completed an informed consent process before commencement of the study.

**Study participant and surgical procedure.** The study participant had complete, non-spastic quadriplegia subsequent to traumatic cervical spine injury sustained 4 years previously while diving. He had full bilateral elbow flexion (grade 5/5), active wrist extension with radial deviation through an incomplete range of motion against gravity (grade 2/5), but no motor function below the level of C6. His sensory level was C5 on the right (because of altered but present light touch on his thumb) and C6 on the left. He had intact proprioception in the right upper limb at the shoulder for internal rotation through external rotation, at the elbow for flexion through extension, at the forearm for pronation through supination, and at the wrist for flexion through extension. Proprioception for right digit flexion through extension at the metacarpal-phalangeal joints was impaired for all digits. His injury was complete, with an overall neurological level of C5 American Spinal Injury Association Impairment Scale A with zone of partial preservation for motor function to C6 bilaterally according to the International Standards for Neurological Classification of Spinal Cord Injury<sup>23</sup>. Electromyogram with nerve conduction studies were performed by a board certified electromyographer to further phenotype the patient's motor impairments, documenting presence of voluntary motor units in the right deltoid (C56), biceps (C56), and extensor carpi radialis (C67), but absence of motor units in triceps (C678), pronator teres (C67), extensor digitorum (C78), extensor indicis (C78), flexor pollicis longus (C78), abductor pollicis brevis (C8–T1), and first dorsal interosseus (C8–T1) muscles.

The patient underwent a left frontoparietal craniotomy. The hand area of motor cortex was identified preoperatively by fusing fMRI activation maps obtained while the patient attempted movements co-registered to the preoperative planning MRI. An intraoperative navigation system was used to plan the craniotomy. The microelectrode array was then implanted into the cortex using a pneumatic inserter. The pedestal was tunnelled under the skin to a posterior exit point. Reference wires were placed subdurally according to the manufacturer's guidelines.

**fMRI acquisition.** A 3 T MRI system (Philips Achieva) equipped with an eight-channel head coil was used to collect ten runs of functional images with a blipped single-shot gradient-echo EPI imaging sequence with a spatial resolution of  $2.2\text{ mm} \times 2.2\text{ mm} \times 2.3\text{ mm}$ . The acquisition parameters were repetition time/echo time ( $T_R/T_E$ ) 2,000/30 ms,  $80^\circ$  flip angle,  $104 \times 75$  matrix size,  $230\text{ mm} \times 168\text{ mm}$  field of view, 27 slices, and echo train of 75 ms. The number of dynamic acquisitions was 165 and each run took 5.5 min. The field of view of the functional scans did not cover the whole brain, but did cover the superior motor strip bilaterally. A high-resolution T1-weighted image was also acquired with the following parameters:  $T_R/T_E$  7.7/3.5 ms,  $1\text{ mm} \times 1\text{ mm} \times 1\text{ mm}$  voxel resolution, three-dimensional acquisition with 144 slices.

We asked the participant to visualize movements of his right hand while we collected fMRI. We collected ten blocks, each with ten trials. Each trial consisted of a fixation cross displayed for 1.4 s, followed by a 5 s video depicting a movement beginning after a 0.1 s delay. Each movement video was repeated three times with a 1.5 s interval between repetitions. The orientation cross for the next set of videos was displayed after a jittered inter-trial interval of between 8 and 12 s, as well as two sets of three repetitions of 5 s videos depicting a right hand at rest. Each video began with a right hand in the resting position of his right hand. The hand in the video then executed a specific motion and held that movement before returning to the rest position. The videos depicted the following eight movements: clenching hand into a fist, wrist flexion, wrist pronation and supination, drumming fingers by sequentially touching the palm with the fifth to second fingers, splaying all fingers to maximum extension, flexing the first finger, flexing the second finger, flexing the third finger; and finally resting in which the hand did not move.

To obtain the final results, we contrasted all of the movement conditions against the rest condition and found that the areas with the highest activation were in the left motor and sensory cortex. The location of the microelectrode array implantation overlapped with an area in the motor strip where activity in response to motor visualization was observed.

**Functional MRI processing.** Functional MRI processing was performed using AFNI<sup>24</sup> and FSL<sup>25</sup>. We dropped the first three volumes of each run, then removed spikes with 3dDespike and regressed out cardiac and respiratory signals with 3dTroicor.

Next, we temporally shifted all of the slices in each volume to the start of the volume's collection using 3dTshift with Fourier interpolation and registered all of the functional volumes to the third volume of the first run. We fitted a general linear model restricted maximum likelihood with regressors for each of the visualization movement types and regressors for motion and the derivative of motion. The regressor for each visualization movement was a 5 s block convolved with a haemodynamic response function starting at the onset of each video presentation. Regressions were fitted with 3dDeconvolve and 3dREMLfit. To identify activity related to motor visualization, we contrasted all of the movements versus the rest condition and applied a threshold of a  $t$  value of 7, then dilated these results so that they were visible on the cortical surface. We co-registered the functional images with the T1-weighted anatomical image with FSL's grey-matter/white-matter boundary-based linear registration. Postoperative computed tomography images were linearly registered with the same anatomical image using the FSL image registration tool with a mutual information cost function and 6 degrees of freedom. The overlap between microelectrode array location and fMRI results was visualized with MRICroGL (<http://www.cabiatl.com/mricrogl/>). Researchers were not blinded during pre-processing or subsequent analyses.

**System architecture.** The Neuroport neural data acquisition system and the Utah microelectrode array with 1.5 mm electrodes (Blackrock Microsystems) were used to acquire the neural data. A 0.3 Hz first-order high-pass and a 7.5 kHz third-order low-pass Butterworth analogue hardware filter were applied to the data, and each of the 96 channels of the microelectrode array were sampled at a rate of 30,000 samples per second. The digitized data were then transmitted to a personal computer where they were decoded to determine which motion was being imagined and then encoded to evoke the desired response from the muscles in the forearm. The computer communicated with the custom high-definition NMES that drove the electrode sleeve wrapped around the forearm.

**Signal processing.** The signal processing and decoding/control algorithms were all run on a personal computer using MATLAB (version 2012a). The digitized data from the Neuroport was processed every 100 ms. Stimulation artefact in the data was detected by looking for a threshold crossing at  $500\text{ }\mu\text{V}$  that occurred simultaneously on at least 4 of 12 randomly selected channels. A 2.5 ms window of data around each detected artefact was then removed and adjacent data segments were rejoined. This approach removed a large portion of the artefact, but stimulation effects still remained. Residual artefact, caused by the Neuroport amplifiers recovering from a brief period of saturation after each stimulation pulse, led to increased signal amplitude during stimulation (see wavelet processing description below). The decoders were trained with data from the period before and during stimulation to accommodate this residual artefact and allow the user to initiate and sustain the correct movement.

Wavelet decomposition was then applied to the data, using the 'db4' wavelet and 11 wavelet scales<sup>26</sup>. Wavelet decomposition has been shown to be an effective tool in neural decoding applications and provides information encompassing single unit, multiunit, and LFP, without requiring spike sorting<sup>27</sup>.

In this study four wavelet scales (3–6) were used, corresponding to the multiunit frequency band spanning approximately 234–3,750 Hz. The mean of the wavelet coefficients for each scale of each channel was calculated every 100 ms and a 1 s wide boxcar filter was applied to smooth the data. Baseline drift in the data was estimated by using a 15 s boxcar filter and was subtracted from the smoothed mean wavelet coefficients for the corresponding 100 ms window. This drift may be due to multiple causes including changes in concentration/focus level and has been observed previously<sup>28</sup>. The mean coefficients of scales 3–6 were then standardized per channel, per scale, by subtracting the mean and dividing by the standard deviation of those scales and channels during the training blocks. The four scales were then combined by averaging the standardized coefficients for each channel, resulting in 96 values for every 100 ms of data. When present, residual artefact in a portion of this 100 ms segment can cause a local increase in wavelet coefficients leading to a twofold to eightfold increase in the reported mean for the segment. The resulting values were then used as features, termed mean wavelet power (MWP), as input into the decoders. Example decoder output scores are shown in Extended Data Fig. 5.

To look at MWP signal quality over the study period, data for all channels were collected over 60 s at the beginning of each test session where the participant was instructed to close his eyes and rest. MWP features with no mean subtraction were calculated to approximate the power in the multiunit frequency bands. After an initial decline, the MWP stabilized after 150 days after implantation (Extended Data Fig. 6).

**Stimulation.** The participant received intermittent stimulation during sessions (with a maximum of three sessions per week, typically lasting 3–4 h including setup time). The continuous, variable output score of each decoder was used to drive the stimulation level during all tasks. Stimulation hydrogel material (Axelgaard) cut into disks, 12 mm in diameter, were applied to the stimulation electrodes. The

centre-to-centre spacing of the electrodes was 22 mm along the long axis of the forearm and 15 mm in the transverse direction. Electrical stimulation was provided intermittently in the form of current-controlled, monophasic (biphasic recommended for long-term, continuous use) rectangular pulses of a 50 Hz pulse rate and 500  $\mu$ s pulse width. Pulse amplitudes ranged from 0 to 20 mA and were updated every 100 ms.

For each motion, the stimulation intensity and spatial pattern were determined by using a trial-and-error method. The output score of a given decoder was in the  $-1$  to  $1$  range and when exceeding zero, the system enabled stimulation for that movement. If the output scores of multiple decoders exceeded zero simultaneously, then the system enabled the movement with the highest decoder score. If 'graded stimulation intensity' was not being used, then the stimulator would use the full (calibrated) stimulation intensity for the desired motion.

In some cases, graded stimulation intensity was used, which required three stimulation intensity levels to be captured. The stimulation intensity levels were set so that they produced a low, medium, and high amount of joint deflection (visually determined by the operator). The decoder output score,  $r$ , was then used to set the stimulation intensity level,  $f(r)$ , during real-time decoding. A piecewise linear function was used to interpolate between the three calibrated stimulation intensities as follows:

$$f(r) = \begin{cases} 0, & r \leq 0 \\ r(m-l)/0.2 + l, & 0 < r \leq 0.2 \\ r(h-m)/0.2 + 2m - h, & 0.2 < r \leq 0.4 \\ h, & r > 0.4 \end{cases}$$

where  $l$ ,  $m$ , and  $h$ , correspond to the three stimulation intensities. The thresholds of 0.2 and 0.4 were derived empirically and were used to compensate for the nonlinear response of the muscles to surface stimulation.

Stimulation spatial patterns and intensity levels were saved in a database. In subsequent sessions with the participant, the previous calibrations could be recalled. See Supplementary Table 1 for electrode labels and total number of stimulation electrodes used to evoke each movement during the individual movement task and Extended Data Fig. 7 for a representative stimulation pattern used to evoke 'hand open' movement.

**Task methods and algorithm training.** The decoders were trained in blocks, each consisting of repetitions of desired motions with short rest periods between. The participant was cued by a small graphical hand displayed on a monitor. After each block, the decoders were adapted (created or updated) using data from the most recent three blocks. During the first block (before a decoder had been built), the system provided 'scripted' stimulation (stimulation corresponding to the cue) while the participant imagined/attempted the cued movements. After the first block, a decoder was built and the study participant received real-time, cortically controlled stimulation. Every 100 ms, the system updated the stimulation with the electrode pattern for the motion with the highest decoder output score. For a short amount of time during each cue, referred to as 'no-feedback time,' no stimulation was given which provided training data from both the period before motion and during stimulation/motion. For tasks that included residual shoulder and elbow movement, longer movement cues were used to also capture training data for the grasp alone and subsequently the grasp with arm movement in space. This training data (cues and MWP features) were used to train a nonlinear support vector machine algorithm, resulting in more robust decoders. A decoder for each motion (against all other motions/rest) was built using a nonlinear Gaussian radial basis function kernel<sup>29</sup> to process the features. The processed features were then input into a non-smooth support vector machine algorithm that used sparsity optimization to improve decoder accuracy by zeroing out the least valuable MWP features<sup>15</sup>. During decoding, all decoders ran simultaneously and the decoder with the highest output score above zero was used to drive the stimulator.

**Individual movement task.** During the individual movement task training, the participant was cued by the graphical hand for all six wrist and finger motions. The decoders were trained in seven successive blocks, each consisting of three trials of each movement. Each cue lasted 2.5 s and was always followed by a rest period of random length (2.5–4.0 s); the movement cues were randomly shuffled. The 'no-feedback time' was set to 1.2 s in the first block and 0.8 s in subsequent blocks. The test portion of the 'individual movement task' used the same cueing parameters as during training, except each movement cue was repeated five times and the 'no-feedback time' was disabled. For this task, graded stimulation intensity was not used. Overall accuracy for the neural decoder was calculated by the fraction of 100 ms segments where the neural decoder agreed with the cue.

To quantify finger and hand movements in the individual movement task, coloured finger cots (coverings) were placed on the participant's fingers. For detecting movements involving the wrist, an additional cot was placed on a plastic cylinder extending out past the participant's thumb. A Bumblebee2 stereo camera was positioned above the participant's hand to track movement in three dimensions. The colour of the cots was used to identify individual fingers and locate them in three-dimensional space using a combination of custom code and OpenCV<sup>30</sup>. The location of the fingers was then fed into the same decoding algorithm used for neural decoding to determine the position of the hand, termed the video classifier. This classifier was trained using video data collected at 12 frames per second using scripted stimulation for the moves of interest. As previously described, accuracy, sensitivity, and specificity were computed. A permutation test was performed by comparing the observed data to data with shuffled labels<sup>16</sup>. To account for reaction and system lag time, the video decoder data were shifted by 1.2 s.

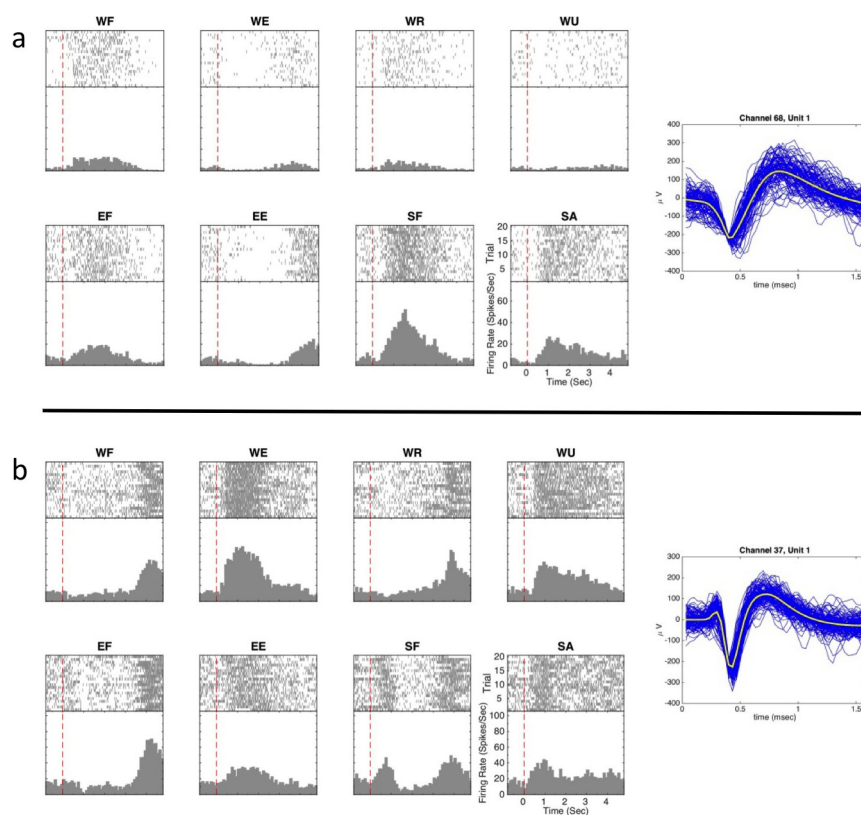
**GRASSP assessment.** Independent decoders were trained separately for each GRASSP test that required stimulation. During the GRASSP training, the participant was cued by the graphical hand for the motion needed for that particular task. The duration of the cue and rest periods were randomly selected from a uniform distribution across the ranges specified in Supplementary Table 2, which also specified the number of repetitions of each cue. Each cue was always followed by a rest period. The 'no-feedback time' was set to 0.8 s for all blocks. The test portion of the task used the same cueing parameters as those used during training, except each movement cue was repeated five times and the 'no-feedback time' was disabled. The GRASSP tests were conducted without any cues. Since decoder sets were trained for each individual GRASSP test (versus training one decoder set for all tests), the user could not switch between GRASSP tests at will without reloading or training the appropriate decoders. Some GRASSP tests required two movements to be trained for one test to open the hand around the object and then grip the object. In some GRASSP tests, graded stimulation intensity was used to make the gripping action smoother.

**Functional movement task.** While performing training for the functional movement task, the participant was cued by the graphical hand for all three of the motions requiring stimulation. Each cue lasted 5.5 s and was always followed by a 5 s rest period. Each movement cue was repeated four times. The movement cues were sequenced so that the participant could perform the complete functional movement task during training. The 'no-feedback time' was set to 1.2 s in the first block and 0.8 s in subsequent blocks. For the hand open and bottle grip movements, the stimulation intensity was not graded; however, graded stimulation intensity was used for the stir bar grip to enable a smooth grasp.

The test portion of the functional movement task was conducted without any cues. The participant was verbally instructed when to start upon which a timer was initiated. To be considered a successful trial, the participant had to grip the bottle, pour at least ten dice into the jar, set the bottle down, pick up a stir stick with a pinch grasp, stir the contents with at least two (circular) stirs, and leave the stir stick in the jar (at which time the timer was stopped). No objects could be dropped or knocked over during the task and the trial had to be completed within 60 s. He was given a total of 10 min to complete as many trials as possible (with a mandatory 60 s rest period between each trial). Sixty seconds of rest was chosen to give the researcher time to reset the props of the task and to give the participant time to rest physically and mentally. Because of his injury, the participant had limited ability to lift his arm, and repetitive tasks such as this requiring him to do so could cause his shoulder to fatigue and become uncomfortable.

23. Kirshblum, S. C. *et al.* International Standards for Neurological Classification of Spinal Cord Injury: cases with classification challenges. *J. Spinal Cord Med.* **37**, 120–127 (2014).
24. Cox, R. W. AFNI: software for analysis and visualization of functional magnetic resonance neuroimages. *Comput. Biomed. Res.* **29**, 162–173 (1996).
25. Smith, S. M. *et al.* Advances in functional and structural MR image analysis and implementation as FSL. *Neuroimage* **23** (Suppl. 1), S208–S219 (2004).
26. Mallat, S. *A Wavelet Tour of Signal Processing* (Academic, 1998).
27. Sharma, G. *et al.* Time stability of multi-unit, single-unit and LFP neuronal signals in chronically implanted brain electrodes. *Bioelectron. Med.* **2**, 63–71 (2015).
28. Kennedy, P. Changes in emotional state modulate neuronal firing rates of human speech motor cortex: a case study in long-term recording. *Neurocase* **17**, 381–393 (2011).
29. Scholkopf, B. *et al.* Comparing support vector machines with Gaussian kernels to radial basis function classifiers. *IEEE Trans. Signal Process.* **45**, 2758–2765 (1997).
30. Bradski, G. The OpenCV library. *Doctor Dobbs J.* **25**, 120–126 (2000).

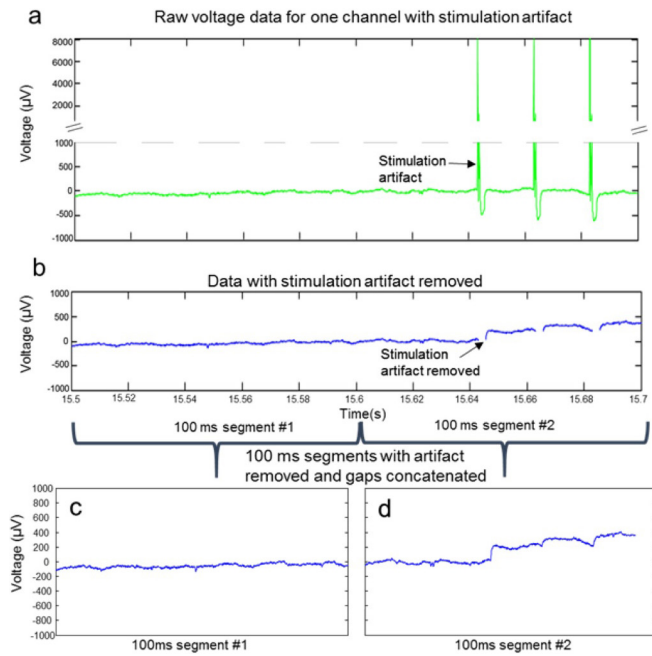




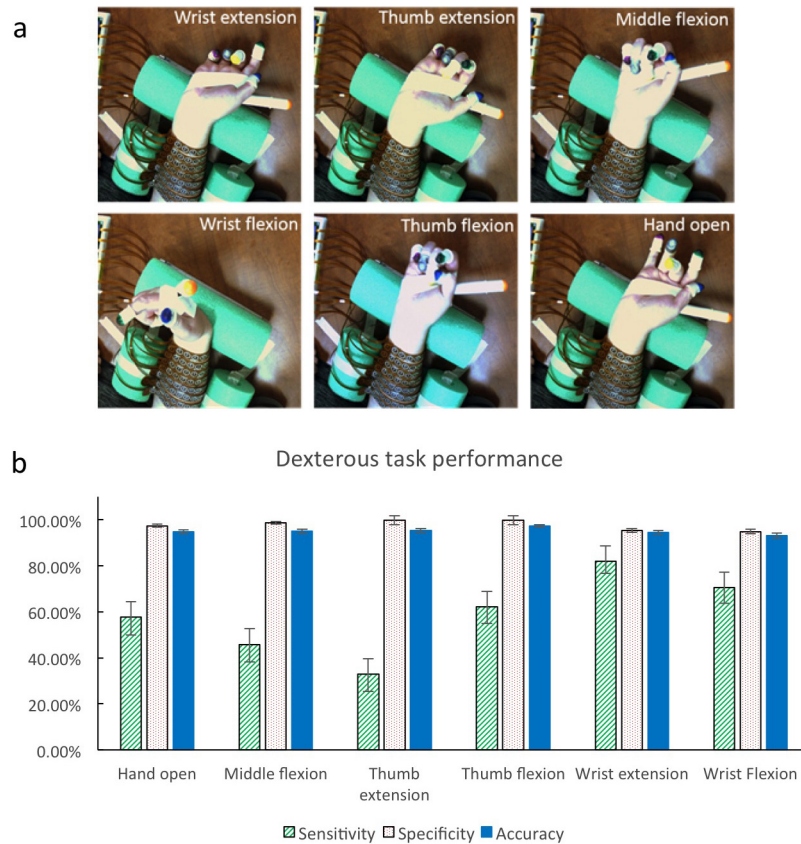
**Extended Data Figure 1 | Neural modulation. a, b,** Examples of rasters and peristimulus histograms from simultaneously recorded units with response to attempted or performed wrist, elbow, and shoulder movements are shown in (a) for channel 68, unit 1 and in (b) for channel 37, unit 1. The participant was presented with cues to attempt wrist flexion (WF), wrist extension (WE), wrist radial deviation (WR), wrist ulnar deviation (WU), elbow extension (EE), or to perform elbow flexion (EF), shoulder flexion (SF), or shoulder abduction (SA). Each cue was presented for a duration of 3 s with a random jitter of 0–2 s followed by a 3 s period with a jitter of 0–2 s. We presented 20 trials of each in random order. The top part of each subpanel is a raster, the black dots represent spikes, each row of spikes represents data from one trial. All trials were aligned on cue

presentation (time zero, red dashed line). On the right of each set of panels is an example of 500 randomly selected waveforms from the discriminated unit (blue). The thick yellow line represents the average waveform for the unit. The top panel shows unit 1 on channel 68. This unit responded well to movement around the shoulder and elbow. The bottom panel shows activity from unit 1 on channel 37 that responded to wrist, elbow, and shoulder movements. Because the participant was asked to imagine and hold the movement throughout the cue period, the latent increase of activity after the cue ended (during the rest period) was probably due to the participant imagining the antagonist movement to return to a neutral position.



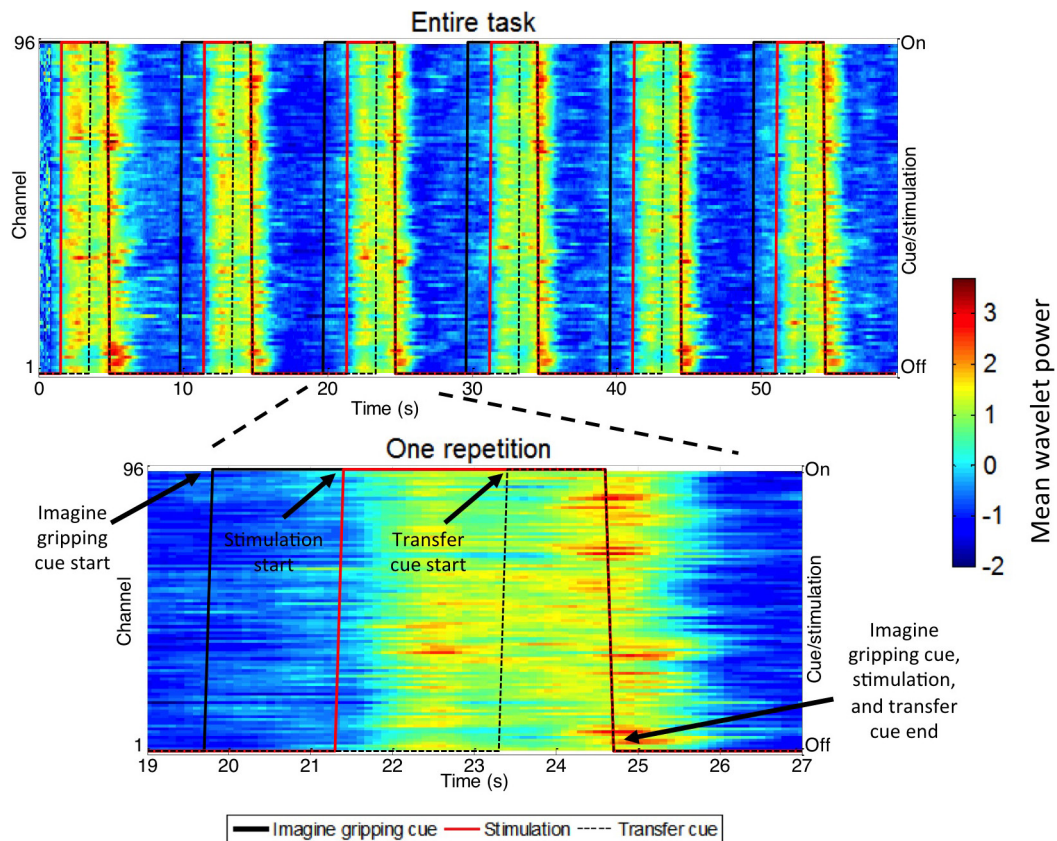


**Extended Data Figure 2 | Stimulation artefact removal.** Stimulation artefact was removed from the signal associated with each electrode before further processing (see Methods). **a**, Example of the voltage data during stimulation for a single electrode shows a period before stimulation and a period during stimulation containing three stimulation artefacts (large amplitude pulses). **b**, Large artefacts were detected and removed to reduce their effects; however, residual artefacts could remain (causing a twofold to eightfold increase in MWP during stimulation). **c**, The period before stimulation is shown for reference. **d**, The period with three stimulation artefacts (after removal) is shown after concatenation/rejoining of the signal segments, and wavelet decomposition is then performed on these shorter, concatenated data (see Methods).



**Extended Data Figure 3 | Individual movement task performance.** The participant was visually cued to attempt each of the six trained movements. **a**, Snapshot of each movement. **b**, The performance (sensitivity, specificity, and accuracy) was measured by automatic evaluation of video frames of hand movements for each of the six moves. The overall accuracy of the NBS was  $70.4 \pm 1.0\%$  ( $P < 0.01$ ). Statistics for individual movements are calculated against a movement specific cue vector with two classes: one class for the movement of interest and one class for all other moves plus rest, which we refer to as the non-target class. Sensitivity is the percentage of video frames during cued movement where the correct movement was observed. It captures the ability of the participant to initiate the specific movement in response to the cue and his ability to sustain the movement for the duration of the two second cue. Specificity is the percentage of video

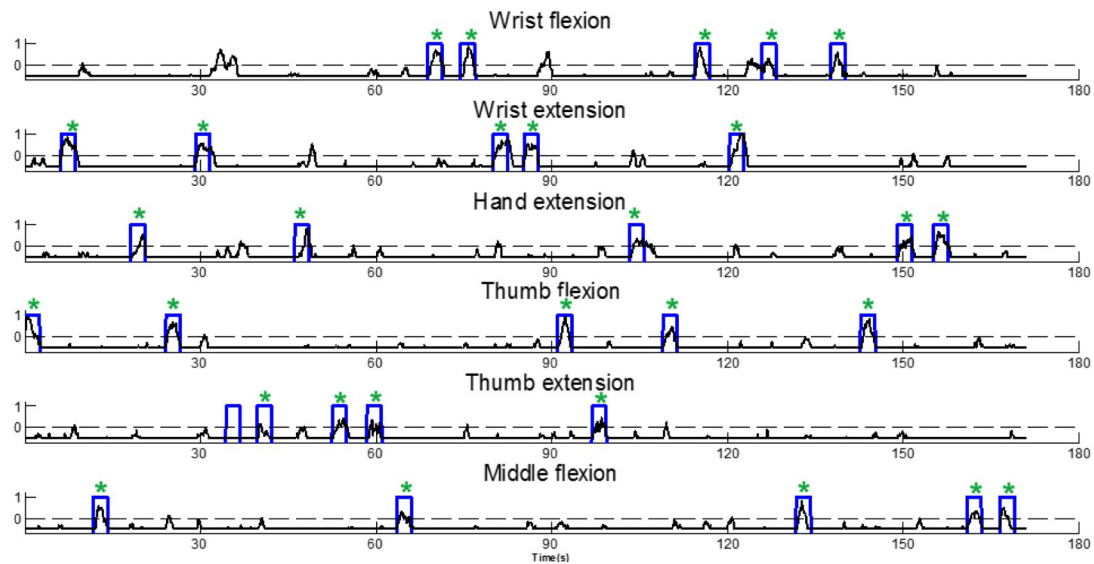
frames during non-target cues that were correctly identified. Accuracy is the percentage of video frames where the observed hand position matched the cue vector for that specific movement. It can also be calculated as a weighted average of sensitivity and specificity, with the weights corresponding to the relative frequency of the two classes. As such, the sensitivity and specificity give additional performance information which can be especially useful when the relative frequency of the classes is skewed towards one class. Overall accuracy measures the performance of all six decoders simultaneously by calculating the percentage of video frames where the observed hand position matches the position the participant was asked to achieve. Thus, errors on any of the six movements are incorporated in the overall accuracy, so it is expected to be lower than the accuracy of any individual movement. Error bars,  $\pm 2$  s.d. (Video taken by D. Friedenberg.)



**Extended Data Figure 4 | MWP during a combination movement task.** MWP as calculated by the NBS algorithm during a task combining imagined movement, stimulation-induced movement, and non-paralysed muscle movement is shown. The participant's reaction time and a 1 s boxcar filter used to smooth the neural data creates the delay observed after the cues are presented. The participant's hand was placed on top of a spoon set on a table and he was cued by the graphical hand to imagine gripping a spoon (solid black line), provided stimulation to evoke the actual gripping of the spoon (red line), and cued by an audible beep to transfer the spoon using the residual movement in his shoulder and elbow

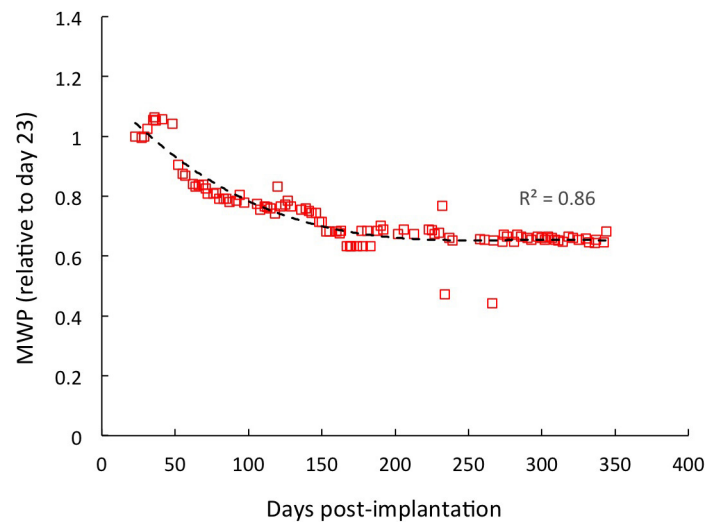
(black dashed line). The spoon transfer distance was approximately 45 cm to the left/right, alternating between each trial. Differences in MWP can be seen for each of the three portions of each trial. The increase in MWP between imagined movement and stimulation-induced movement is caused by residual stimulation artefact. During the transfer cue, the utilization of residual (shoulder) movement was associated with a different and consistent MWP pattern, requiring the development of a more robust decoding strategy which combined data from imagined, stimulation-induced, and transfer movements for decoder training (see Methods).





**Extended Data Figure 5 | Neural decoder outputs for each movement in the individual movement task.** The blue line represents the cue that the participant was trying to match. For a particular movement, when the blue line was at one, the user was prompted to imagine that movement. The black line is the decoder output for that move. The decoded movement is rest if none of the decoder outputs are above zero (dotted line); otherwise,

it is determined by the maximum of all the decoder outputs. Green stars are placed above the 29 out of 30 cues in which the neural decoder correctly matched the cue. In this plot the cues have been shifted by 0.8 s to account for reaction and system lag time. Decoder output below  $-0.5$  has been set to  $-0.5$  for visual clarity.

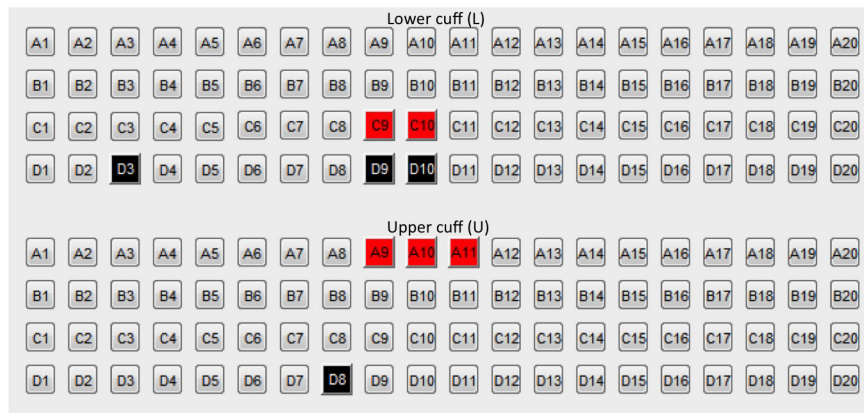


**Extended Data Figure 6 | Evolution of MWP over time.** At the beginning of each test session, data for all channels were collected over 60 s where the participant was instructed to close his eyes and rest. No stimulation was provided during this period. MWP features with no mean subtraction

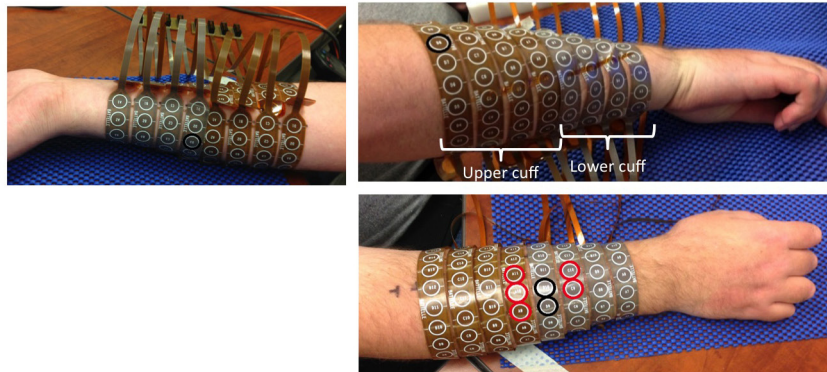
were calculated to approximate the power in the multiunit wavelet bands corresponding to scales 3–6. After an initial decline the MWP stabilized at 150 days after implantation. Dashed line represents a polynomial fit of order 4.

a

## Stimulation electrodes for “Hand open”



b



**Extended Data Figure 7 | Stimulation electrode pattern.** The stimulator was calibrated to evoke movements in the hand and wrist. **a**, Representative map of the anode (red) and cathode (black) electrodes in the lower (L) and upper (U) stimulation cuffs used to evoke the ‘hand

open’ movement during the individual movement task. **b**, Stimulation cuffs on the participant’s arm, with the stimulation pattern highlighted. See Supplementary Table 1 for a complete list of electrode patterns used for the individual hand movement task. (Photographs taken by N. Annetta.)

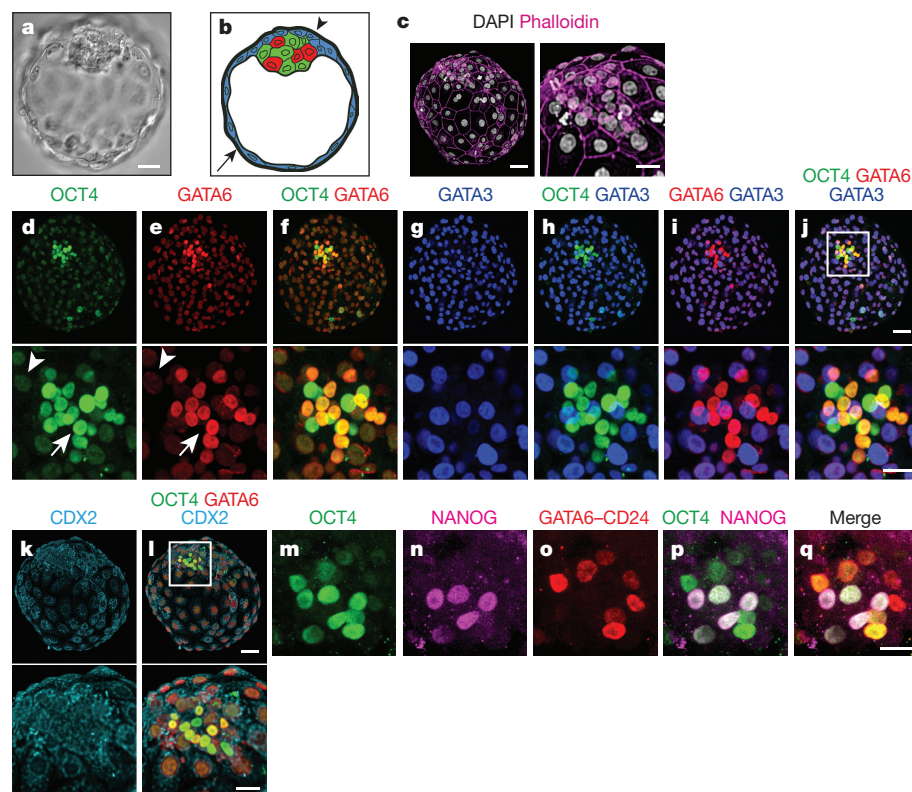
# Self-organization of the *in vitro* attached human embryo

Alessia Deglincerti<sup>1\*</sup>, Gist F. Croft<sup>1\*</sup>, Lauren N. Pietila<sup>1</sup>, Magdalena Zernicka-Goetz<sup>2</sup>, Eric D. Siggia<sup>3</sup> & Ali H. Brivanlou<sup>1</sup>

Implantation of the blastocyst is a developmental milestone in mammalian embryonic development. At this time, a coordinated program of lineage diversification, cell-fate specification, and morphogenetic movements establishes the generation of extra-embryonic tissues and the embryo proper, and determines the conditions for successful pregnancy and gastrulation. Despite its basic and clinical importance, this process remains mysterious in humans. Here we report the use of a novel *in vitro* system<sup>1,2</sup> to study the post-implantation development of the human embryo. We unveil the self-organizing abilities and autonomy of *in vitro* attached human embryos. We find human-specific molecular signatures of early cell lineage, timing, and architecture. Embryos display key landmarks of normal development, including epiblast expansion, lineage segregation, bi-laminar disc formation, amniotic and yolk sac cavitation, and trophoblast diversification. Our findings highlight the species-specificity of these developmental events and provide a new understanding of early human embryonic development beyond the blastocyst stage. In addition, our study establishes a new model system relevant to early human pregnancy

loss. Finally, our work will also assist in the rational design of differentiation protocols of human embryonic stem cells to specific cell types for disease modelling and cell replacement therapy.

Recently established *in vitro* implantation platforms in the mouse have expanded our knowledge of post-implantation development and recapitulated *in vivo* early developmental landmarks<sup>1,2</sup>. However, extrapolation from mouse to human is limited since mammalian embryos display species-specific differences in post-implantation morphology and many molecular markers have not been validated<sup>3–5</sup>. Therefore, to gain insights into the self-organizing abilities of human embryos, we sought to establish an *in vitro* attachment platform. To determine the dynamics of cell fate specification, we monitored the morphology and the expression of cell-type-specific markers in pre- and post-attachment embryos: the inner cell mass (ICM) and epiblast (Epi) markers OCT4 and NANOG, the ICM and primitive endoderm (PE) marker GATA6, and the trophoblast (TE) markers CDX2 and GATA3 (ref. 6). Additionally, to distinguish TE subtypes, we used cytokeratin 7 (CK7) and  $\beta$ -human chorionic gonadotropin (HCGB)<sup>7</sup>. Embryos were fixed and stained every other day starting at day

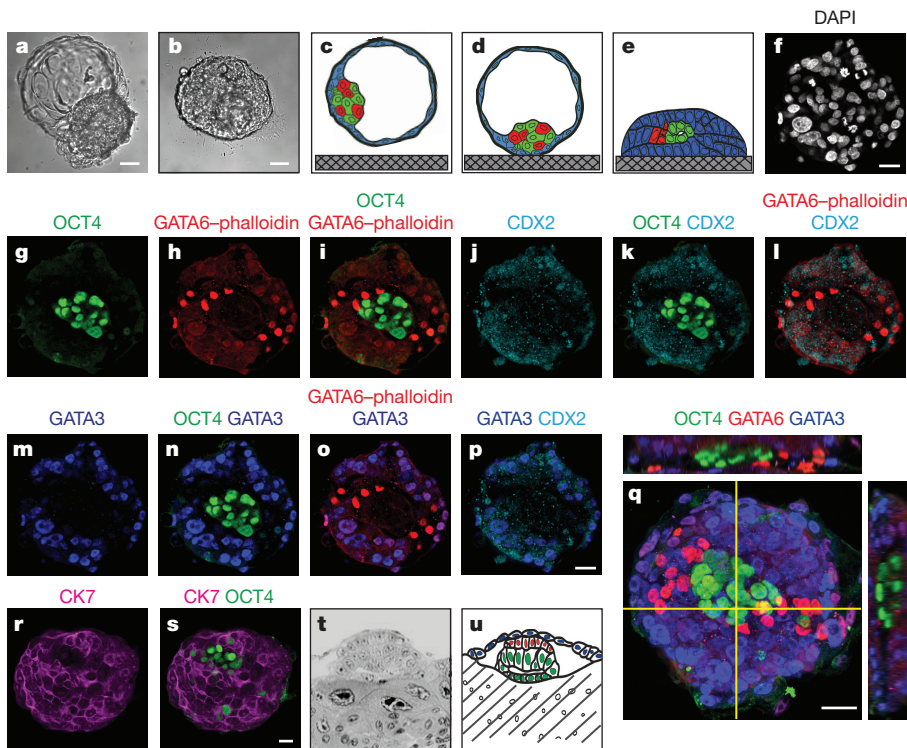


**Figure 1 | D.p.f. 6 human blastocyst embryos display human-specific transcriptional profiles.** **a**, DIC image of a d.p.f. 6 human blastocyst; scale bar, 100  $\mu$ m. **b**, Cartoon of a d.p.f. 6 embryo with salt-and-pepper distribution of OCT4 (green) and OCT4/GATA6 (red) ICM cells, and GATA3 (blue); mural (arrow) and polar (arrowhead) TE cells. **c–q**, Immunostaining of d.p.f. 6 blastocysts. **c**, Three-dimensional rendering of the front half of a d.p.f. 6 blastocyst (left) and ICM-zoom (right). DAPI (white) and phalloidin (magenta, actin, virtual channel). **d–j**, Three-dimensional rendering of a d.p.f. 6 blastocyst (top) and ICM-zoom (bottom, box in **j**) stained for OCT4 (green), GATA6 (red, virtual channel), GATA3 (blue) ( $n = 3–8$ ); arrows indicate high marker levels in ICM, arrowhead low levels in TE. **k–l**, Whole embryo (top) and ICM-zoom (bottom, box in **l**) of the d.p.f. 6 blastocyst from **c**; OCT4 (green), GATA6 (red, virtual channel), CDX2 (cyan) ( $n = 5$ ). **m–q**, ICM-zoom of d.p.f. 6; OCT4 (green), NANOG (magenta), GATA6 and CD24 (red,  $n = 3$ ). Scale bar, 100  $\mu$ m for whole embryos, 20  $\mu$ m for ICM-zooms.

<sup>1</sup>Laboratory of Stem Cell Biology and Molecular Embryology, The Rockefeller University, New York, New York 10065, USA. <sup>2</sup>Department of Physiology, Development, and Neuroscience, University of Cambridge, Physiology Building, Downing Street, Cambridge CB2 3DY, UK. <sup>3</sup>Center for Studies in Physics and Biology, The Rockefeller University, New York, New York 10065, USA.

\*These authors contributed equally to this work.



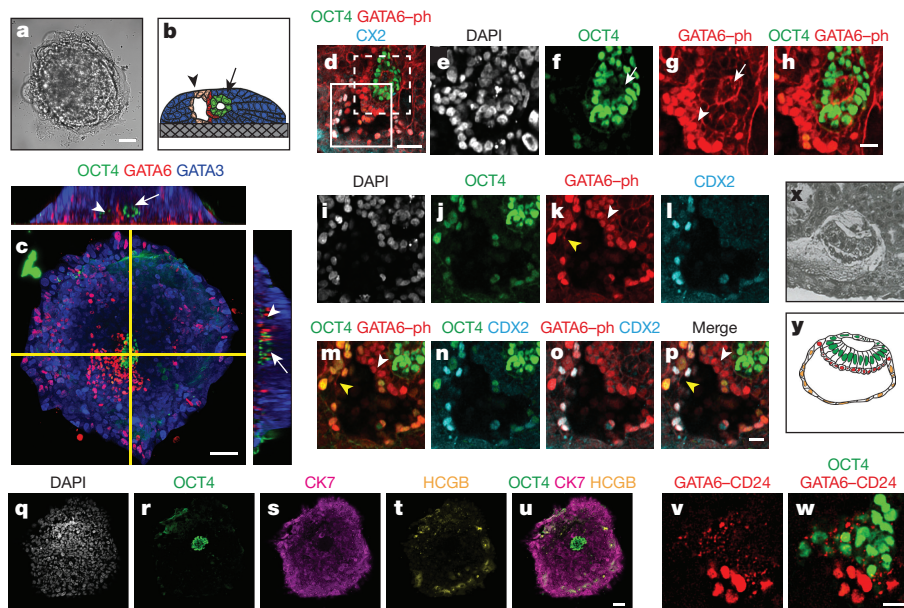


**Figure 2 | Attached d.p.f. 8 human embryos begin transcriptional and morphological self-organization.** **a, b**, DIC images of **(a)** attaching d.p.f. 7.5 embryos with the TE (top left) collapsing onto the embryo and **(b)** attached and compacted d.p.f. 8. Scale bar, 50  $\mu$ m. **c–e**, Cartoons of blastocyst **(c)** floating, **(d)** orienting to attach on the side of the polar TE, and **(e)** attached; Epi (green), PE (red), TE (blue). **f–p**, Optical sections of d.p.f. 8 embryos stained with the indicated markers ( $n = 4$ ). Scale bar, 20  $\mu$ m. **q**, Three-dimensional rendering of the embryo in **f–p** with OCT4 (green), GATA6 (red), and GATA3 (blue), flanked by xz (top) and yz (side) views at coordinates indicated by yellow lines. Scale bar, 50  $\mu$ m. **r, s**, Optical sections of d.p.f. 8 embryo stained with the indicated markers. Scale bar, 20  $\mu$ m. **t, u**, Carnegie stage 5a section **(t)** and cartoon **(u)**; Epi (green), PE (red).

post-fertilization (d.p.f.) 6, and counterstained with 4',6-diamidino-2-phenylindole (DAPI) to detect and count nuclei and phalloidin to delineate cell boundaries (Supplementary Video 1).

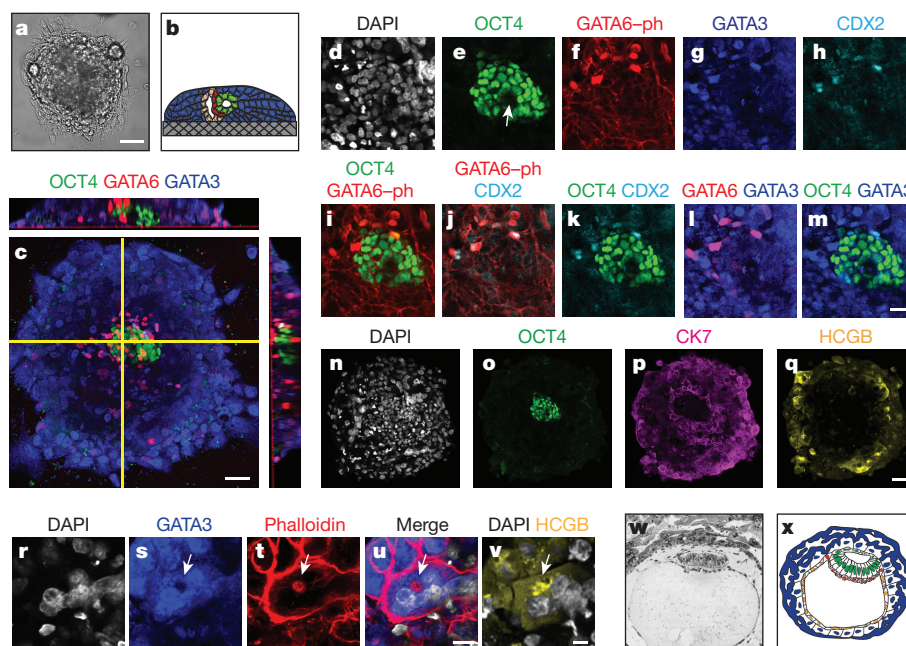
The pre-implantation human blastocyst (d.p.f. 6–7) is a hollow ball of cells composed of tightly packed, small nuclei in the ICM

and larger cells in the TE ( $n = 8$ ; Fig. 1a–c, Extended Data Figs 1a–c and 2a, Supplementary Table and Supplementary Videos 2 and 3). All ICM cells showed high intensity OCT4 staining (OCT4<sup>HI</sup>), with a subset having high GATA6 (GATA6<sup>HI</sup>) in a salt-and-pepper manner (Fig. 1d–f and Extended Data Fig. 2b–d). NANOG was observed



**Figure 3 | D.p.f. 10 embryo cell lineages diversify and self-organize amniotic and yolk sac cavities.** **a, b**, DIC image of a d.p.f. 10 embryo (scale bar, 50  $\mu$ m) **(a)** and cartoon of a section through the embryo **(b)**; Epi (green), amniotic cavity (arrow), PE (red), yolk sac cavity (arrowhead), ysTE (orange), TE (blue). **c**, Three-dimensional rendering of a d.p.f. 10 embryo stained with OCT4 (green), GATA6 (red), and GATA3 (blue) ( $n = 4$ ), flanked by xz (top) and yz (side) views at coordinates indicated by the yellow lines, with amniotic (arrow) and yolk sac cavities (arrowhead). Scale bar, 50  $\mu$ m. **d**, Optical section of the central region from **c**; OCT4 (green), GATA6-phalloidin (red), and CDX2 (cyan) ( $n = 4$ ). Scale bar, 20  $\mu$ m. **e–h**, Optical section of the dashed box in **d**, but on a z-plane

6  $\mu$ m higher, with the amniotic cavity at the centre (arrow) lined by PE (arrowhead). Scale bar, 20  $\mu$ m. **i–p**, Optical section of the box in **d**, but on a z-plane 4  $\mu$ m below with the yolk sac cavity at the centre lined by PE (white arrowhead) and ysTE (yellow arrowhead). Scale bar, 20  $\mu$ m; **q–u**, Optical section of a d.p.f. 10 embryo stained with the indicated markers ( $n = 4$ ). Scale bar, 100  $\mu$ m. **v, w**, Optical section of the Epi/PE area of a d.p.f. 10 embryo stained with OCT4 (green) and GATA6-CD24 (red) ( $n = 4$ ). Scale bar, 20  $\mu$ m. **x, y**, Carnegie stage 5b section **(x)** and cartoon **(y)** with Epi (green), amniotic cavity, PE (red), yolk sac cavity, and ysTE (orange). ph, phalloidin.



**Figure 4 | D.p.f. 12 embryos exhibit characteristic of CS5C *in vivo* including TE cellular phenotypes.** **a, b**, DIC image of a d.p.f. 12 embryo (**a**) (scale bar, 50  $\mu$ m) and cartoon of a side view (**b**) with Epi (green), amniotic and yolk sac cavities, PE (red), ysTE (orange), and TE (blue). **c**, Three-dimensional rendering of a d.p.f. 12 embryo stained for OCT4 (green), GATA6 (red), and GATA3 (blue) ( $n = 2$ ); flanked  $xz$  (top) and  $yz$  (side) views at indicated coordinates. **d–m**, Epi zoom of **c** stained with the markers indicated; the arrow in **e** points at the amniotic cavity.

Scale bar, 20  $\mu$ m. **n–q**, Optical section of the embryo in **c** re-stained with the indicated markers ( $n = 2$ ); scale bar, 100  $\mu$ m. **r–v**, Appearance of ST phenotypes; multinucleated ST cells, DAPI (white), GATA3 (blue), phalloidin (red), HCGB (yellow), and nascent lacuna (arrow). Scale bar, 20  $\mu$ m. **w–x**, Carnegie stage 5c section (**w**), and cartoon (**x**) of Epi (green), amniotic and yolk sac cavities, PE (red), ysTE (orange), CT, ST, and lacunae in TE (blue).

exclusively in OCT4<sup>+</sup> ICM cells, which did not stain for GATA6 (Fig. 1m–q). No GATA3 or CDX2 were detected in the ICM. All TE cell nuclei stained for GATA3, as well as low intensity OCT4 (OCT4<sup>LO</sup>), and GATA6 (GATA6<sup>LO</sup>; Fig. 1g–j) but not NANOG. Variable levels of CDX2 were also detected in TE (CDX2<sup>+</sup>; Fig. 1k–l and Extended Data Fig. 2e, f); however, in most cases, surprisingly, the signal was localized to the cytoplasm instead of the nucleus (Fig. 1k–l). No HCGB and weak, non-filamentous CK7 were detected (Extended Data Fig. 2g;  $n = 6$ ). The ubiquitous OCT4 pattern we report is in agreement with previous staining of human blastocysts at d.p.f. 5–7 (refs 6, 8, 9), and stands in contrast to the mouse where it is ICM specific. GATA6 expression in the ICM is different from the mouse and previous reports in humans<sup>6,8,9</sup>. In the mouse, the salt-and-pepper ICM distribution becomes restricted to PE and physically sorted from the Epi population by late blastocyst<sup>10</sup>. Our data suggest that physical sorting between Epi and PE has not yet occurred in human d.p.f. 6, even if the transcriptional profile is consistent with specification of presumptive Epi and PE fate. While both CDX2 and GATA3 are TE specific, we find that GATA3 is a better marker of human TE because of its high signal and consistent nuclear localization. Our results also confirm the molecular profiles from single-cell RNA sequencing studies of laser-dissected blastocyst (d.p.f. 5–6)<sup>11</sup>, and expand them by providing new spatial resolution. Thus, at d.p.f. 6, we define two cell populations in the ICM: OCT4<sup>HI</sup>/NANOG<sup>+</sup>/GATA6<sup>−</sup> (presumptive Epi precursors), and OCT4<sup>HI</sup>/GATA6<sup>HI</sup>/NANOG<sup>−</sup> (presumptive PE precursors); we also define one population of TE cells: OCT4<sup>LO</sup>/GATA6<sup>LO</sup>/GATA3<sup>+</sup>/CDX2<sup>+</sup>.

We next allowed blastocysts to attach using a recently established mouse *in vitro* implantation platform<sup>1,2</sup>. We found that most human blastocysts (70%,  $n = 91$ ) attached between d.p.f. 7 and 8 (Fig. 2a, b). Interestingly, despite their random original orientation, by d.p.f. 7 the blastocysts always attached on the side of the polar TE, the portion of TE that is the closest to the ICM (Fig. 2c–e). At d.p.f. 8, all embryos adopted a flattened structure (Fig. 2b, e, f, Extended Data

Fig. 3a, Supplementary Table and Supplementary Video 4). While the total cell number remained constant, the relative number of cells contributing to different territories changed (Extended Data Fig. 1a–c and Supplementary Table). Two discrete territories emerged from the ICM, and formed the Epi and PE lineages, respectively characterized by OCT4<sup>HI</sup>/GATA6<sup>−</sup> and OCT4<sup>−</sup>/GATA6<sup>HI</sup> (Fig. 2g–i). Epi cells formed a tight core at the centre of the attached embryos (Fig. 2g). Phalloidin staining was extremely faint in all embryos in two separate experiments, suggesting qualitatively different actin status in the immediate post-attachment period. In the TE, OCT4 and GATA6 staining completely disappeared. CDX2 remained at very low and variable levels and was restricted to weak nuclear staining of a subset of TE cells at the periphery (Fig. 2j–l), while GATA3 demarcated all OCT4<sup>−</sup> and GATA6<sup>−</sup> TE nuclei surrounding Epi and PE (Fig. 2m–q). At this stage, all GATA3<sup>+</sup> cells also showed strong filamentous CK7 staining (Fig. 2r–s), a marker of general TE lineage and cytotrophoblast (CT), suggesting molecular progression of TE differentiation compared with d.p.f. 6; however, no HCGB, a marker of syncytiotrophoblast (ST), was detected (Extended Data Fig. 3b)<sup>7</sup>. Taken together, this evidence suggests that in humans, Epi/PE cell sorting occurs at a later stage than in the mouse, around the time of implantation.

At d.p.f. 10, the attached embryos increased in size while maintaining a flattened morphology (Fig. 3a–c, Extended Data Fig. 1a–c and 3b, Supplementary Table and Supplementary Video 5). The OCT4<sup>HI</sup> Epi at the centre of the embryo was more compacted, and formed a cavity delineated by phalloidin: the putative amniotic cavity (Fig. 3d–h, Extended Data Figs 4a–r and 5 and Supplementary Videos 5–7). To ask if the Epi population had progressed in its differentiation, we stained for CD24, which has been reported to increase in primed (Epi-derived) versus ‘naive’ mouse and human embryonic stem cells (ESCs)<sup>12</sup>. CD24 was selectively observed on the surface of human Epi cells starting only at d.p.f. 10 ( $n = 4$  of 4; Fig. 3v–w and Extended Data Fig. 4s–x), and was not detected at d.p.f. 6 (Fig. 1o;  $n = 3$  of 3) or d.p.f. 8 (Extended



Data Fig. 3c;  $n = 4$  of 4), suggesting a developmental transition. All Epi nuclei were also NANOG<sup>+</sup> (Extended Data Fig. 4s–x). It is therefore the d.p.f. 10 Epi population (OCT4<sup>+</sup>/NANOG<sup>+</sup>/CD24<sup>+</sup>), and not the d.p.f. 6 or d.p.f. 8 Epi, that most closely matches the profile of human ESCs<sup>12</sup>, which supports a recent hypothesis on the developmental origin of human ESCs<sup>9</sup>. GATA6<sup>HI</sup> PE cells were distributed in a layer juxtaposed to one side of the Epi, delineating the bi-laminar germ disc (Fig. 3c–p and Extended Data Fig. 4a–r). On the other side of the PE cells, a second, larger cavity self-organized: the putative yolk sac cavity (Fig. 3i–p, Extended Data Figs 4l–r and 5 and Supplementary Videos 5 and 6). A novel population of CDX2<sup>+</sup>/GATA6<sup>LO</sup>/OCT4<sup>LO</sup> cells, never described before in other mammals, lined the yolk sac cavity in all embryos examined ( $n = 4$ ). While it is possible that these cells represent parietal endoderm, we believe that they are a novel population because of their unique molecular signature. We therefore named these cells ‘yolk sac trophoderm’ or ysTE (Fig. 3p). GATA3<sup>+</sup> cells continued to demarcate the entire TE (Extended Data Fig. 4a–k), which also co-stained for CK7 (Fig. 3q–u). However, a subset of GATA3<sup>+</sup>/CK7<sup>+</sup> cells now stained for HCGB (Fig. 3t), suggesting the emergence of the HCGB<sup>+</sup> ST lineage<sup>7</sup>. Collectively, the architecture of the embryo at this stage closely resembles Carnegie stage 5b–c<sup>13</sup>, with amniotic cavity, yolk sac cavity, bi-laminar disc in between, and TE differentiation (Fig. 3x–y). This demonstrates that the embryo alone can direct both lineage specification and diversification, as well as tissue morphogenesis and architectural organization, without maternal input.

At d.p.f. 12, the amniotic and yolk sac cavities, albeit present, appeared collapsed (Fig. 4a–l, Extended Data Figs 6 and 7a–j, Supplementary Table and Supplementary Video 8). Epi and PE expression profiles were unchanged from d.p.f. 10, while TE showed further differentiation and organization. HCGB staining increased in intensity and clusters expanded to generate multinucleated cells with lacunae, the *in vivo* site of maternal blood vessel invasion, therefore adding functional cytological characteristics of ST lineage progression (Fig. 4n–q and Extended Data Fig. 7k–o). TE subtypes self-organized in concentric rings, with a ring of CT surrounding the Epi, multiple foci of ST towards the edge of the embryo, and a layer of CT adjacent to the substrate (Fig. 4r–v). These features closely mimic and provide the first molecular signature of histological samples<sup>13</sup> (Carnegie stage 5b–c; Fig. 4w–x). We concluded our experiments at d.p.f. 14, in accordance with internationally recognized bioethical guidelines<sup>14,15</sup>. Outgrowths at this late stage lost interpretable relation to *in vivo* correlates<sup>13</sup> ( $n = 8$ ), suggesting the limitations of our two-dimensional culture environment.

In this study we show that by simply providing an attachment substrate, human blastocysts self-organize to recapitulate many key features of *in vivo* development, surprisingly independently of maternal input at least up to d.p.f. 12. Additionally, the differences between human and mouse embryos, including unique architecture, cell types, and tissue organization, emphasize the necessity of working with human embryos to understand human development. By defining the molecular composition of cell lineages in the early embryo, our data provide a new *ex vivo* reference to reinterpret controversial discrepancies between mouse and human stem cells and their derivatives. This platform also constitutes a new model for a variety of poorly understood placental and embryonic disorders and early pregnancy loss. These and future studies on the origin and regulation of early embryonic cell types will lead to more developmentally based and rational approaches to reprogramming, disease modelling, and cell replacement therapies.

**Online Content** Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

**Received 6 November 2015; accepted 30 March 2016.**

**Published online 4 May 2016.**

1. Bedzhov, I., Leung, C. Y., Bialecka, M. & Zernicka-Goetz, M. *In vitro* culture of mouse blastocysts beyond the implantation stages. *Nature Protocols* **9**, 2732–2739 (2014).
2. Bedzhov, I. & Zernicka-Goetz, M. Self-organizing properties of mouse pluripotent cells initiate morphogenesis upon implantation. *Cell* **156**, 1032–1044 (2014).
3. Rossant, J. Mouse and human blastocyst-derived stem cells: vive les differences. *Development* **142**, 9–12 (2015).
4. Rossant, J., Chazaud, C. & Yamanaka, Y. Lineage allocation and asymmetries in the early mouse embryo. *Phil. Trans. R. Soc. Lond. B* **358**, 1341–1349 (2003).
5. Berg, D. K. et al. Trophoderm lineage determination in cattle. *Dev. Cell* **20**, 244–256 (2011).
6. Niakan, K. K. & Eggan, K. Analysis of human embryos from zygote to blastocyst reveals distinct gene expression patterns relative to the mouse. *Dev. Biol.* **375**, 54–64 (2013).
7. Li, Y. et al. BMP4-directed trophoblast differentiation of human embryonic stem cells is mediated through a DeltaNp63+ cytotrophoblast stem cell state. *Development* **140**, 3965–3976 (2013).
8. Roode, M. et al. Human hypoblast formation is not dependent on FGF signalling. *Dev. Biol.* **361**, 358–363 (2012).
9. O’Leary, T. et al. Tracking the progression of the human inner cell mass during embryonic stem cell derivation. *Nature Biotechnol.* **30**, 278–282 (2012).
10. Schrode, N., Saiz, N., Di Talia, S. & Hadjantonakis, A. K. GATA6 levels modulate primitive endoderm cell fate choice and timing in the mouse blastocyst. *Dev. Cell* **29**, 454–467 (2014).
11. Blakeley, P. et al. Defining the three cell lineages of the human blastocyst by single-cell RNA-seq. *Development* **142**, 3151–3165 (2015).
12. Shakiba, N. et al. CD24 tracks divergent pluripotent states in mouse and human cells. *Nature Commun.* **6**, 7329 (2015).
13. O’Rahilly, R. & Müller, F. Developmental stages in human embryos: revised and new measurements. *Cells Tissues Organs* **192**, 73–84 (2010).
14. National Research Council Human Embryonic Stem Cell Research Advisory Committee, Board on Life Sciences, Board on Health Sciences Policy, National Research Council, and Institute of Medicine. *Final Report of the National Academies’ Human Embryonic Stem Cell Research Advisory Committee and 2010 Amendments to the National Academies’ Guidelines for Human Embryonic Stem Cell Research* (National Academies Press, 2010).
15. International Society for Stem Cell Research. *Guidelines for the Conduct of Human Embryonic Stem Cell Research, Version 1* (ISSCR, Northbrook, Illinois, 2006).

**Supplementary Information** is available in the online version of the paper.

**Acknowledgements** We thank the members of the Brivanlou laboratory for their advice and criticisms, in particular C. Nchako, S. Tse for technical assistance, and members of the Zernicka-Goetz laboratory for their advice on how to culture embryos through attachment. We also thank A.K. Hadjantonakis for discussions, A. Wilkerson for support, and A. Brivanlou and P. Carleton-Evans for their comments on the manuscript. This work was supported by a STARR Foundation grant (number 2013-026) and Rockefeller Private funds. Images were obtained using instrumentation in The Rockefeller University Bio-Imaging Resource Center purchased with grant funds from the Sohn Conference Foundation. The Carnegie stage images are used with permission from the Virtual Human Embryo Project (<http://virtualhumanembryo.rockefeller.edu>). We give special thanks for technical advice on imaging to A. North, K. Thomas, and P. Ariel, and on image analysis and rendering to T. Tong. This work would not have been possible without the generosity of the people who consented to donate their embryos to research, to whom we are indebted.

**Author Contributions** A.D., G.C., and L.P. performed experiments; A.D. and G.C. analysed experiments; M.Z.-G. was instrumental in teaching and transferring knowledge on the mouse technology to A.D.; E.S. provided criticism of the work and manuscript; A.H.B. conceived and designed the project, established contact with the source of the biological material, provided guidance and advice throughout the work, and interfaced with the Institutional Review Board at The Rockefeller University; all authors contributed to the manuscript.

**Author Information** Reprints and permissions information is available at [www.nature.com/reprints](http://www.nature.com/reprints). The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to A.H.B. ([brvnlou@rockefeller.edu](mailto:brvnlou@rockefeller.edu)).

## METHODS

**Ethics statement.** This work was approved by The Rockefeller University Institutional Review Board and was in accord with University policy for research involving human embryos. The informed consent process for embryo donation complied with the NAS Guidelines for Human Embryonic Stem Cell Research and the Tri-Institutional Stem Cell Research Operating Procedures for ESCRO Reviewed Research. The Tri-Institutional Stem Cell Initiative Embryonic Stem Cell Research Oversight (Tri-SCI ESCRO) Committee determined that the use of human embryos for basic developmental studies did not require review and approval by the TriSCI ESCRO Committee because it did not involve derivation of, or research with, human pluripotent stem cells.

**Embryo culturing protocol.** Cryopreserved d.p.f. 5 embryos were thawed using a THAW-KIT1 (Vitrolife) according to the manufacturer's instructions. Briefly, embryos were thawed by placing the cryovials in a 37 °C water bath for about 30–60 s, collected, transferred, and incubated in ETS1 for 5 min at room temperature, 22–23 °C transferred and incubated in ETS2 for 5 min at room temperature, transferred and incubated in ETS3 for 7.5 min at room temperature, transferred and incubated in cryo-PBS for 5 min at room temperature and 5 min at 37 °C (not in the incubator). Embryos were morphologically scored upon thaw and 24 h after culture. To remove the zona, embryos were briefly exposed to acidic Tyrode's solution (Sigma). Embryos were seeded on ibiTreat microscopy  $\mu$  plates (Ibidi), filled with pre-equilibrated IVC1 medium (Cell Guidance Systems). Half of the medium was replaced after 24 and 48 h in culture. Zona-free embryos typically attached 48–60 h after seeding. At this point, the medium was exchanged to IVC2 (Cell Guidance Systems), and embryos were cultured for the indicated times, with half of the medium being replaced with fresh medium every 24 h. The embryo culturing and imaging protocol has been deposited on Nature Protocol Exchange, doi 10.1038/protex.2016.022.

**Immunofluorescence.** Embryos were fixed in 4% PFA in PBS for 30 min at 4 °C, washed three times in PBS, quenched in wash buffer (PBS + 0.1% Triton X-100) plus 100 mM glycine, blocked (wash buffer + 0.1% sodium azide + 10% normal donkey serum), stained for 3–12 h with primary antibodies in blocking buffer, washed, stained with labelled secondary antibodies (Alexa Fluor 488, 555, and 647), DAPI, and phalloidin 647, and imaged (see below). To maximize the molecular information from each embryo, we followed several sequential staining strategies. First, in preliminary stains we found that GATA6 was exclusively nuclear and generated almost no background staining; therefore, to maximize information content from each embryo, we stained all embryos with phalloidin-647 (the same fluorescence channel with GATA6) since these could be segregated during image analysis using a three-dimensional segmentation mask on the DAPI to generate a phalloidin-only or a GATA6-only virtual channel. Floating blastocysts were stained with either DAPI-CDX2-OCT4-GATA6 and phalloidin in the same channel, or DAPI-GATA3-OCT4-GATA6 and phalloidin in the same channel since CDX2 and GATA3 antibodies were of the same species. After analysis and quantification of all attached embryos (d.p.f. 8–14), CDX2 cell numbers were found to be very small and weakly stained (except for the ySTE cells at d.p.f. 10–14); therefore all attached embryos were then re-stained with GATA3 (same species, then same colour (Alexa Fluor 555) as CDX2) and reimaged. The staining for GATA3 was substantially brighter than the previous mouse CDX2-555 signal, except for the

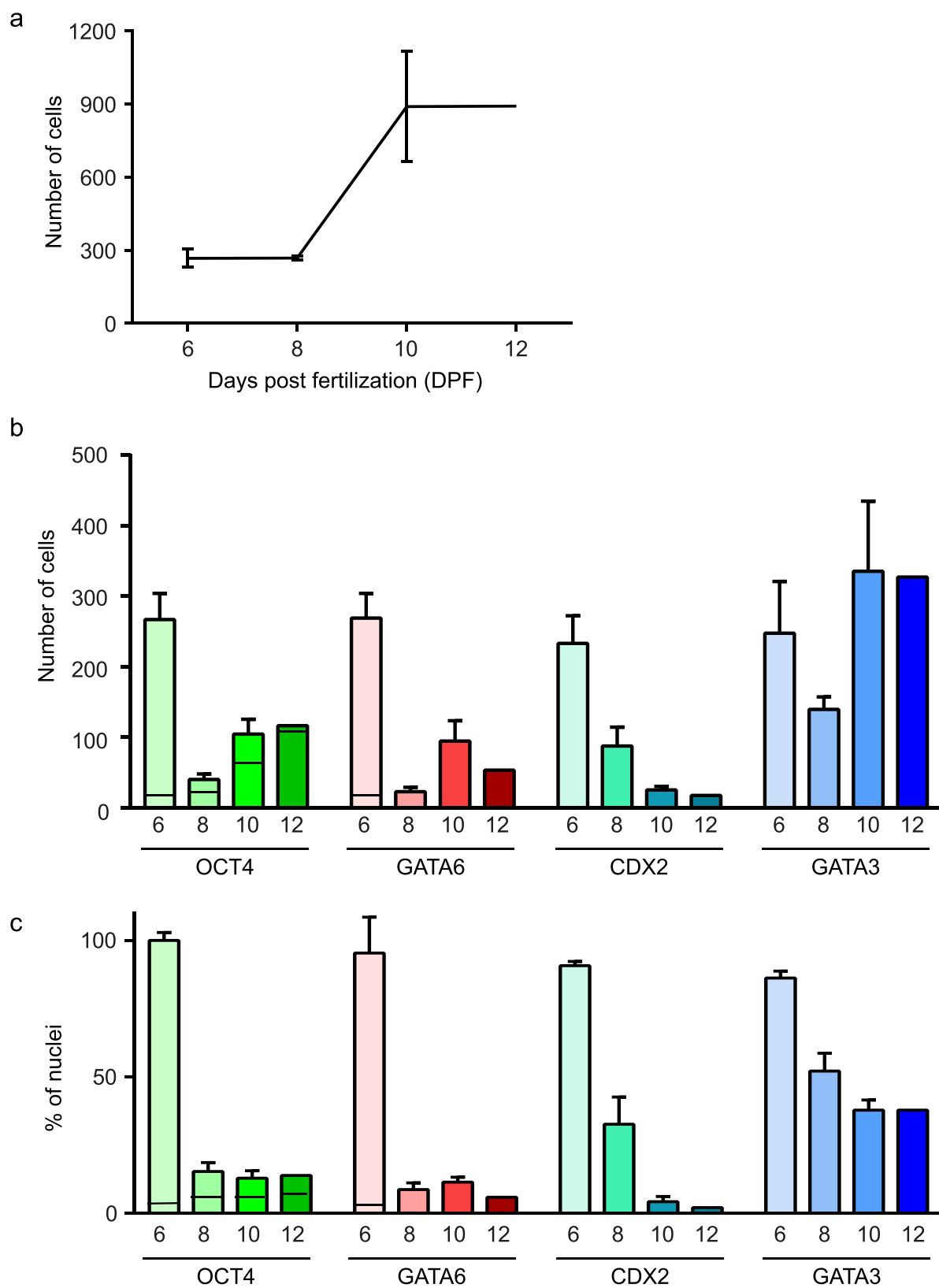
ySTE cells at d.p.f. 10–14. Therefore an extra virtual channel could be superimposed on the same embryos to maximize spatial information, indicating bona fide GATA3 expression in previously CDX2<sup>−</sup> or CDX2<sup>low</sup> TE cells. Similarly, since GATA6 nuclear staining was absent at d.p.f. 14, we re-stained two embryos each with rabbit anti-OTX2 and rabbit anti-SOX2. To further enhance molecular resolution, we then probed TE lineage diversification by staining previously stained attached embryos or fresh day 7 blastocysts. For previously stained attached embryos, we observed that phalloidin completely detached by 2 weeks after staining and was no longer detectable, therefore leaving the 647 channel open except for GATA6<sup>+</sup> nuclei. We therefore blocked embryos with proteolytic Fc fragments (mouse and rabbit), then with unlabelled donkey anti-mouse and -rabbit IgG Fab1, then stained with 647- and 594-labelled donkey anti-mouse and -rabbit IgG Fab1 (all Jackson), and imaged the embryos to measure background and residual crosstalk. We next stained with primary antibodies to trophoblast markers CK7 and HCGB followed by 647- and 594-labelled donkey anti-mouse and -rabbit IgG Fab1 and reimaged. Primary antibodies used were the following: goat-OCT4, Santa Cruz 8628, 1:500; rabbit-GATA6, Cell Signaling 23345, 1:1,000; mouse-NANOG, BD 560482, 1:500; mouse-GATA3, Pierce MA1-028, 1:100; mouse-CDX2, BioGenex CDX2-88, 1:10; rabbit-CK7 Abcam 181598, 1:400; mouse-HCGB, Abcam ab9582, 1:100; mouseCD24-PE conjugate, Abcam 7729, 1:5; rabbit-SOX2, Cell Signaling 35795, 1:200; rabbit-OTX2, Abcam 114138, 1:200. Labelled secondary antibodies, DAPI, and phalloidin-Alexa Fluor 647 (Life) were incubated for 3–12 h and washed before imaging the samples.

**Image acquisition, analysis, and rendering.** Z-stack images were acquired on a Leica SP8 inverted confocal microscope at 12 bits in 1024 pixels  $\times$  1024 pixels using an HCX PL APO CS  $\times$  20/0.75 numerical aperture air-immersion or an HC PL APO CS2  $\times$  40/1.10 numerical aperture water-immersion objective, at 1 Airy unit pinhole diameter and 0.985–1.9  $\mu$ m spacings with manual laser-power Z intensity-compensation, and avoiding detector saturation with laser excitation/HyD detector emission settings, in nanometres, as follows: 405 diode laser 405/415–486, WLL 492/500–550, WLL 552/560–621, and WLL 649/657–710 for DAPI, Alexa Fluor 488, 555, 594, and 647 respectively. When 594 was imaged, the detector setting for Alexa Fluor 555 was adjusted to 560–580, and a fifth track was added, 594 WLL 594/602–631. Images were then deconvolved with a three-dimensional blind algorithm (ten iterations) using AutoDeblur X software (Autoquant). Images were rendered and visualized in Imaris, where total and marker-positive nuclei were quantified by a spot finding algorithm with manual adjustment for fluorescence intensity threshold and quality. Where antibodies or tissue generated non-specific background staining, a nuclear mask was first generated from the DAPI signal, and spots were counted on masked channels as above. Final images were processed and assembled using ImageJ. For the three-dimensional renderings and movies, individual channels were gamma-adjusted to enhance visualization.

**Statistical analysis.** No statistical methods were used to predetermine sample size. The experiments were not randomized. The investigators were not blinded to allocation during experiments and outcome assessment.

Errors and error bars represent the s.e.m. from a minimum of four independent embryos unless otherwise indicated. The figures display representative results. Unless otherwise specified, the results were the same across all the embryos analysed.

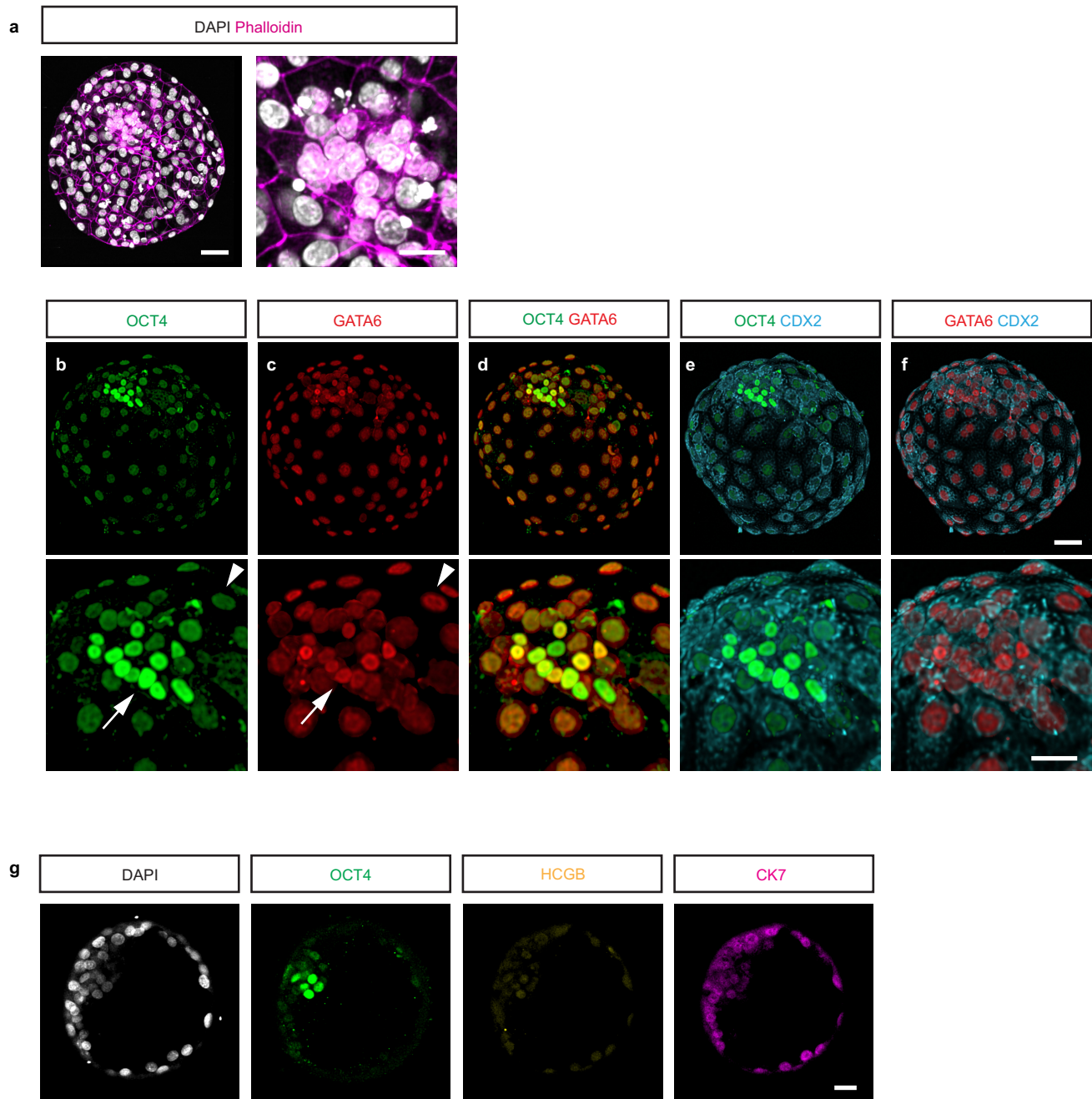




Extended Data Figure 1 | See next page for caption.

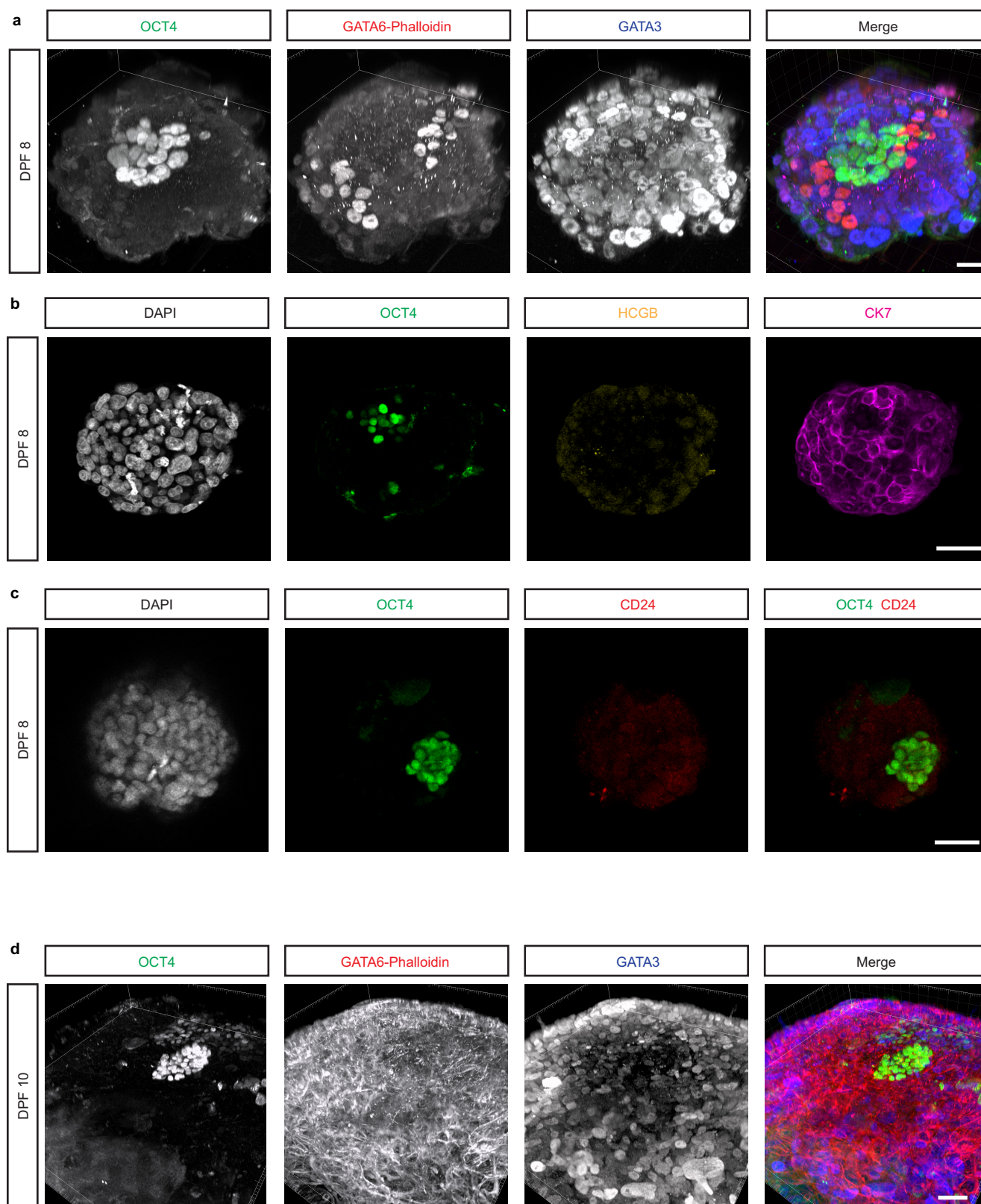
**Extended Data Figure 1 | Quantification of cells positive for transcriptional markers over d.p.f. 6–12 of *in vitro* culture.** **a**, Time course of total cell numbers via DAPI quantification. Confocal data sets for each embryo were scored for total cell numbers in Imaris. The mean  $\pm$  s.e.m. is shown for d.p.f. 6, 8, 10, 12 ( $n=8, 4, 4, 2$ , embryos per time point respectively over  $n=2, 2, 2, 1$  independent experiments). A total of 91 embryos were thawed and 64 (70%) showed developmental progression and attached; 26 of 64 (40%) embryos showed normal size and morphology and were analysed for markers of cell fate. The remaining embryos all showed abnormal development, represented by few (0–5) and scattered OCT4 cells, and were excluded from analysis as obvious failures of development. Subsequent to staining and analysis of this set of embryos with these core markers, six d.p.f. 7 and five d.p.f. 10 embryos showing normal OCT4 development were also stained and analysed for additional markers. **b**, Total cells expressing cell type markers, mean  $\pm$  s.e.m. for OCT4, GATA6, CDX2, and GATA3: d.p.f. 6 ( $n=8, 7, 5, 3$  respectively);

d.p.f. 8 and 10 ( $n=4$  embryos per marker); d.p.f. 12 ( $n=2$  embryos per marker). The line subdividing the OCT4 population identifies the two subpopulation of OCT4<sup>+</sup> cells: OCT4<sup>HI</sup> (bottom of the bar), passing a threshold for high-intensity staining; and OCT4<sup>LO</sup> (top of the bar), passing a low threshold (but not the higher one for OCT4<sup>HI</sup>) for staining. At d.p.f. 6, OCT4<sup>HI</sup> cells were confined to the ICM while the OCT4<sup>LO</sup> cells represented the TE; between d.p.f. 8 and 12, OCT4<sup>LO</sup> cells were also co-expressing CDX2 (ysTE cells), while OCT4<sup>HI</sup> cells were confined to the Epi. GATA6 was scored for high or low (HI/LO) expression at d.p.f. 6 only, where GATA6<sup>HI</sup> cells (bottom of the bar) localized to the ICM and GATA6<sup>LO</sup> cells localized to the TE. See also Supplementary Table. **c**, Number of cells expressing lineage-specific markers expressed as a percentage of total nuclei; each bar shows mean  $\pm$  s.e.m. for OCT4, GATA3, GATA6, and CDX2. The line subdividing the OCT4 bars at d.p.f. 6–12 and the GATA6 bar at d.p.f. 6 represents the high or low (HI/LO) expressing cells as explained above.



**Extended Data Figure 2 | Human-specific transcriptional profiles of ICM and TE in d.p.f. 6 human blastocysts.** **a**, Additional three-dimensional rendering of the front half of a d.p.f. 6 blastocyst (left), and ICM-zoom (right). DAPI (white) identifies all nuclei and phalloidin (magenta, actin, virtual channel) shows intercellular boundaries between TE cells. A three-dimensional segmentation mask on the DAPI channel was subtracted from the raw GATA6–phalloidin channel to yield a phalloidin-only virtual channel for these panels. This embryo is the same as displayed in Fig. 1d–j. Scale bar, 100  $\mu$ m for whole embryos, 20  $\mu$ m for ICM-zooms. **b–f**, Top, whole embryo, and bottom, ICM-zooms, stained for OCT4 (green), GATA6 (red, virtual channel), CDX2 (cyan), and

serial merges. **b**, **c**, Arrows show that ICM cells have high-intensity OCT4 with or without GATA6 staining; arrowheads show low-level expression of OCT4 and GATA6 in TE cells; **e**, **f**, CDX2 staining is weak and predominantly cytoplasmic. The GATA6-only channel is the result of an inverse mask using three-dimensional segmentation on DAPI to remove extra-nuclear phalloidin staining, resulting in nuclear the GATA6-only virtual channel. This embryo is the same as displayed in Fig. 1k–l. Scale bar, 100  $\mu$ m for whole embryos, 20  $\mu$ m for ICM-zooms. **g**, No HCGB and very weak CK7 signal was observed in d.p.f. 7 embryos ( $n = 6$ ). Shown is a section through a d.p.f. 7 embryo stained with DAPI (white), OCT4 (green), HCGB (yellow), and CK7 (magenta). Scale bar, 100  $\mu$ m.



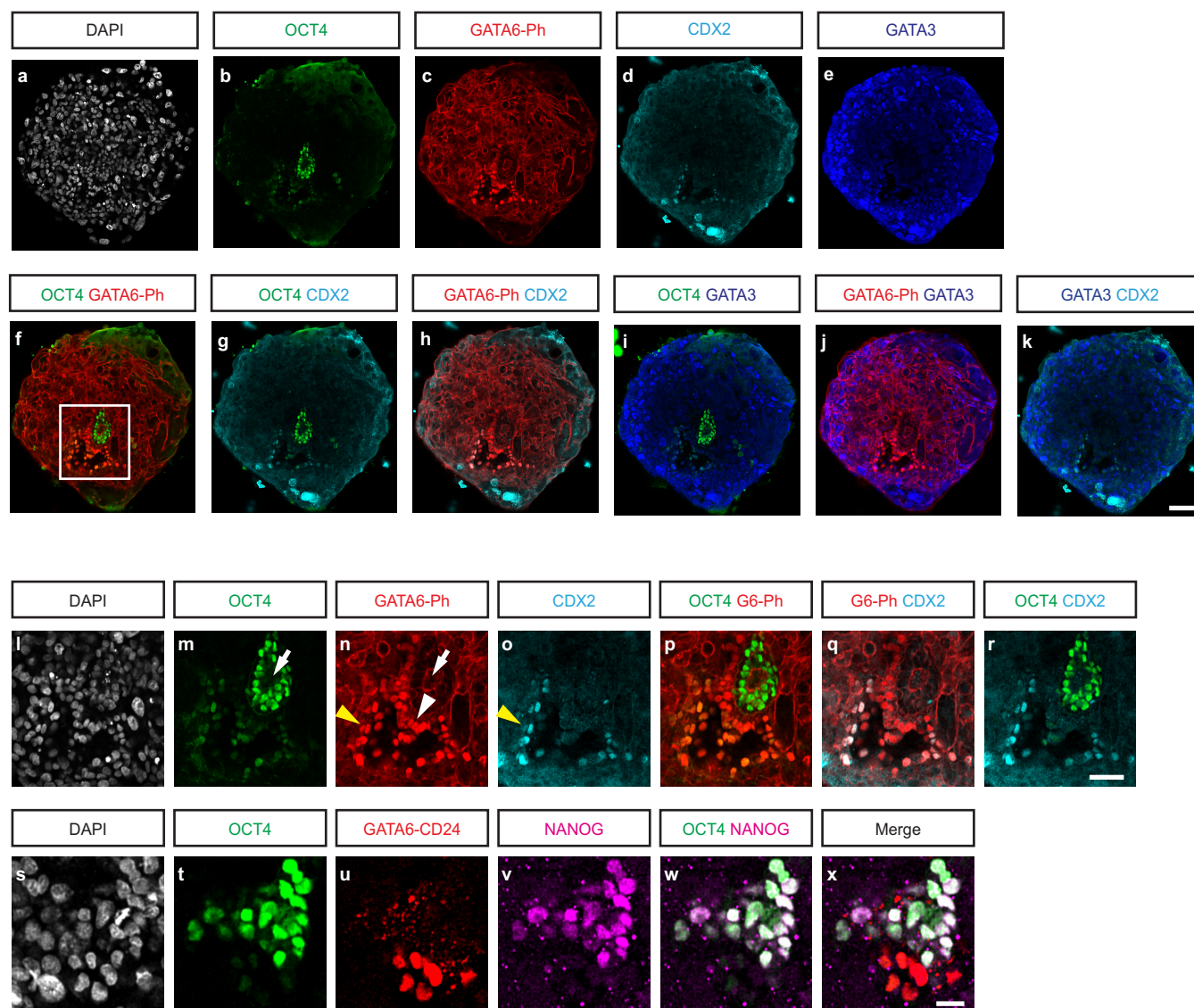
### Extended Data Figure 3 | Molecular signature of d.p.f. 8–10 embryos.

**a**, Embryo reconstruction of a d.p.f. 8 embryo oriented at a 30° pitch; individual channels are shown in greyscale and then merged as OCT4 (green), GATA6-phalloidin (red), GATA3 (blue). Phalloidin staining of all d.p.f. 8 embryos was exceptionally weak compared with all other d.p.f. (all embryos analysed in two separate experiments); scale bar, 50  $\mu$ m.

**b**, No HCGB staining was detected at d.p.f. 8 ( $n = 4$ ). Section through a d.p.f. 8 embryo showing DAPI (white), OCT4 (green), HCGB (yellow),

and CK7 (magenta). Scale bar, 50  $\mu$ m. **c**, No CD24 staining was detected at d.p.f. 8 ( $n = 4$ ). Section through a d.p.f. 8 embryo showing DAPI (white), OCT4 (green), CD24 (red). Scale bar, 50  $\mu$ m. **d**, Embryo reconstruction of a d.p.f. 10 embryo oriented at a 30° pitch; individual channels are shown in greyscale and then merged as OCT4 (green), GATA6-phalloidin (red), GATA3 (blue). Phalloidin staining of this embryo was so intense that the original underlying GATA6 nuclear stain is not visible in this reconstruction; scale bar, 50  $\mu$ m.

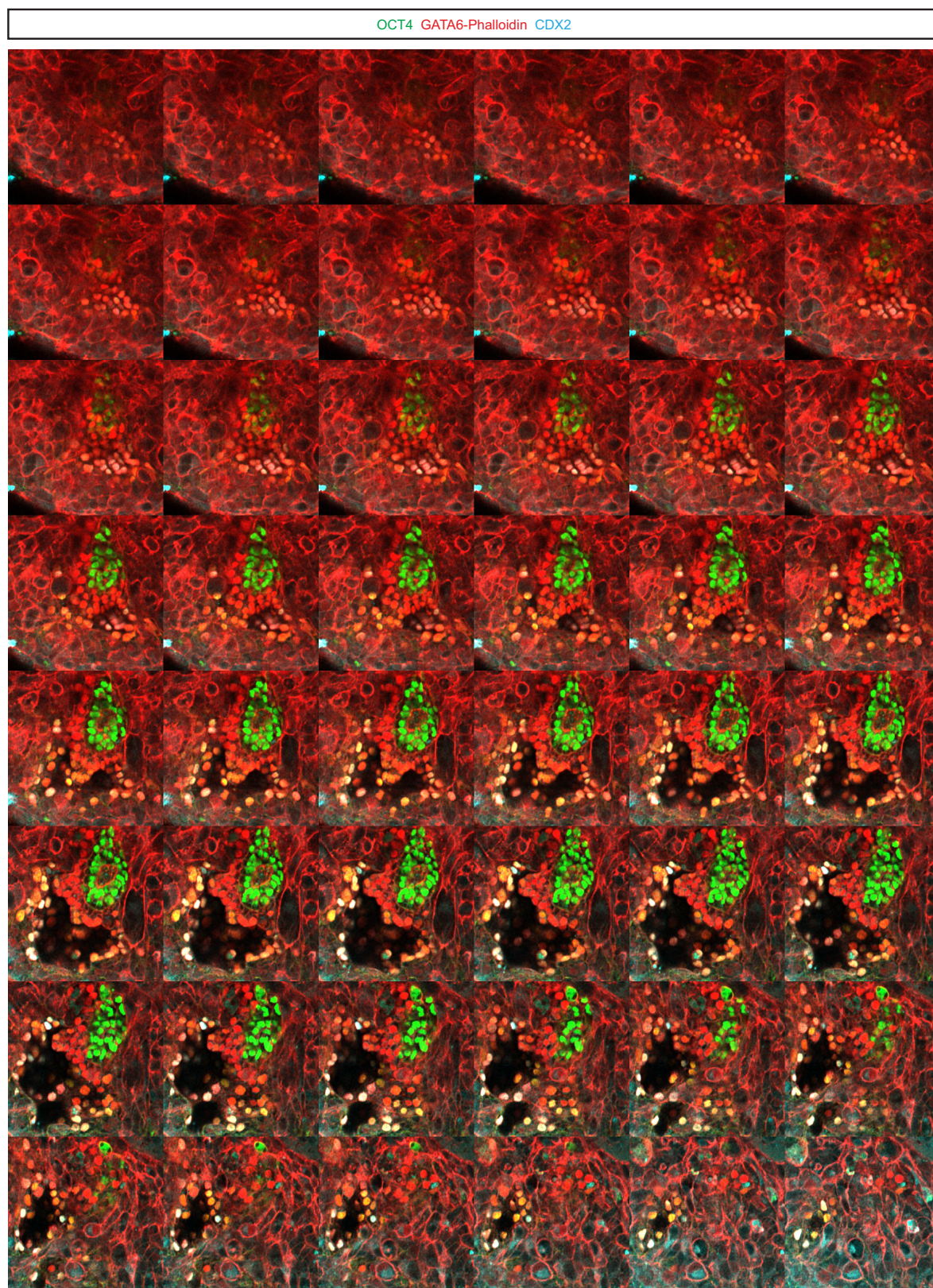




**Extended Data Figure 4 | Human-specific transcriptional profile of d.p.f. 10 embryos.** D.p.f. 10 *in vitro* attached embryos recapitulate the principal landmarks of Carnegie stage 5b (*in vivo* d.p.f. 9). **a–k**, Confocal z-sections of a d.p.f. 10 embryo stained for DAPI (white), OCT4 (green), GATA6 and phalloidin (GATA6-Ph, red), CDX2 (cyan), GATA3 (blue), and combinatorial merges. The box in **f** indicates the area including amniotic and yolk sac cavities, which is shown in Fig. 3d; scale bar, 50  $\mu$ m. **l–r**, Additional confocal z-section of the boxed region in **f**, stained for DAPI (white), OCT4 (green), GATA6–phalloidin (red), CDX2 (cyan), and

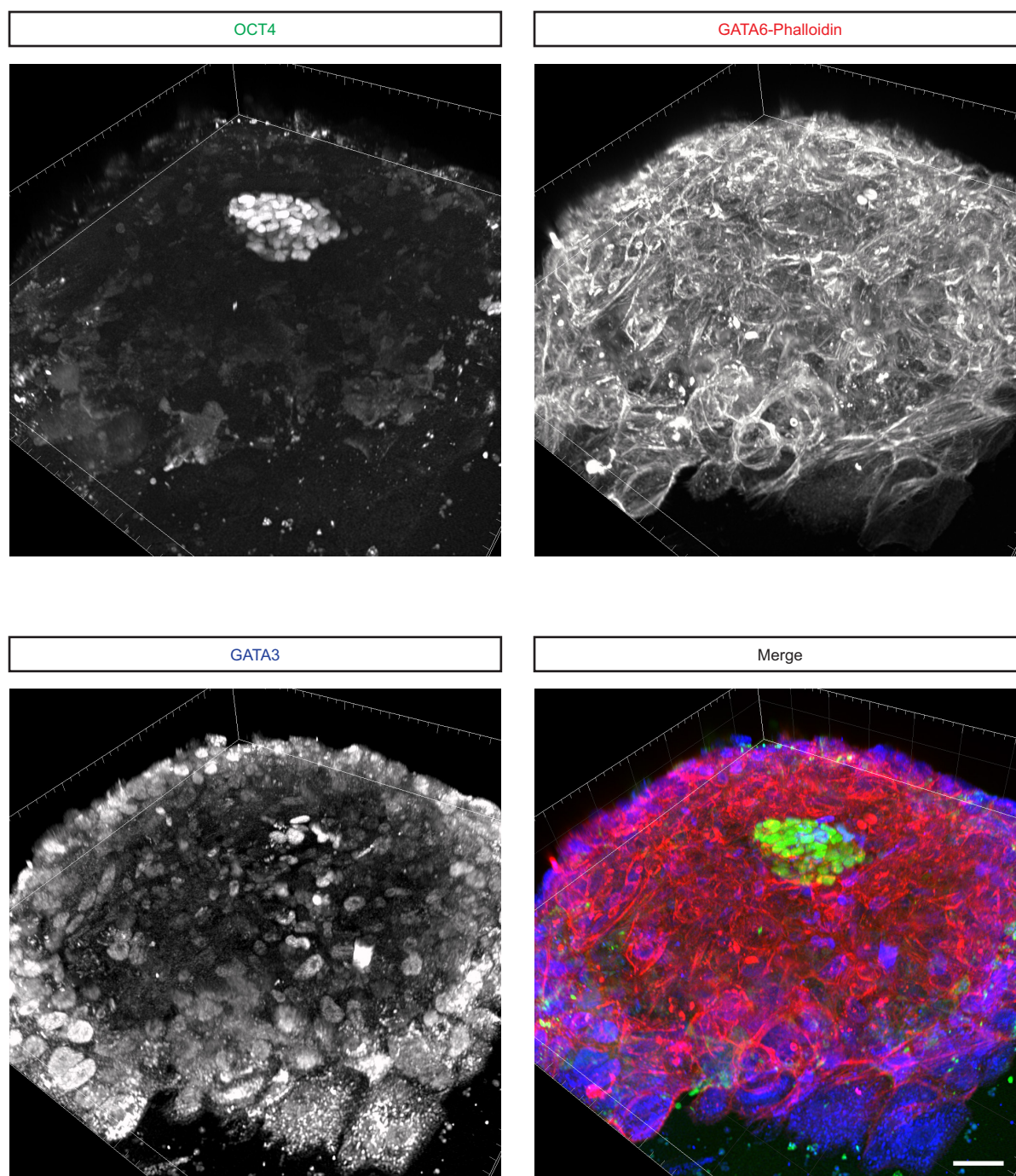
combinatorial merges, showing amniotic cavity formed by OCT4<sup>+</sup>-only Epi cells (arrow in **m** and **n**), flanking GATA6<sup>+</sup>-only PE cells (white arrowhead in **n**), and CDX2<sup>+</sup>/GATA6<sup>+</sup>/OCT4<sup>LO</sup> ysTE cells lining the yolk sac cavity (yellow arrowhead in **n**, **o**); the CDX2<sup>+</sup>/GATA6<sup>+</sup>/OCT4<sup>LO</sup> merge of this series is shown in Fig. 3d. Scale bar, 20  $\mu$ m. **s–x**, Confocal z-section of the Epi/PE area of a d.p.f. 10 embryo stained for DAPI (white), OCT4 (green), GATA6 and CD24 (red), NANOG (magenta). This is the same embryo as in Fig. 3v, w. Scale bar, 20  $\mu$ m.



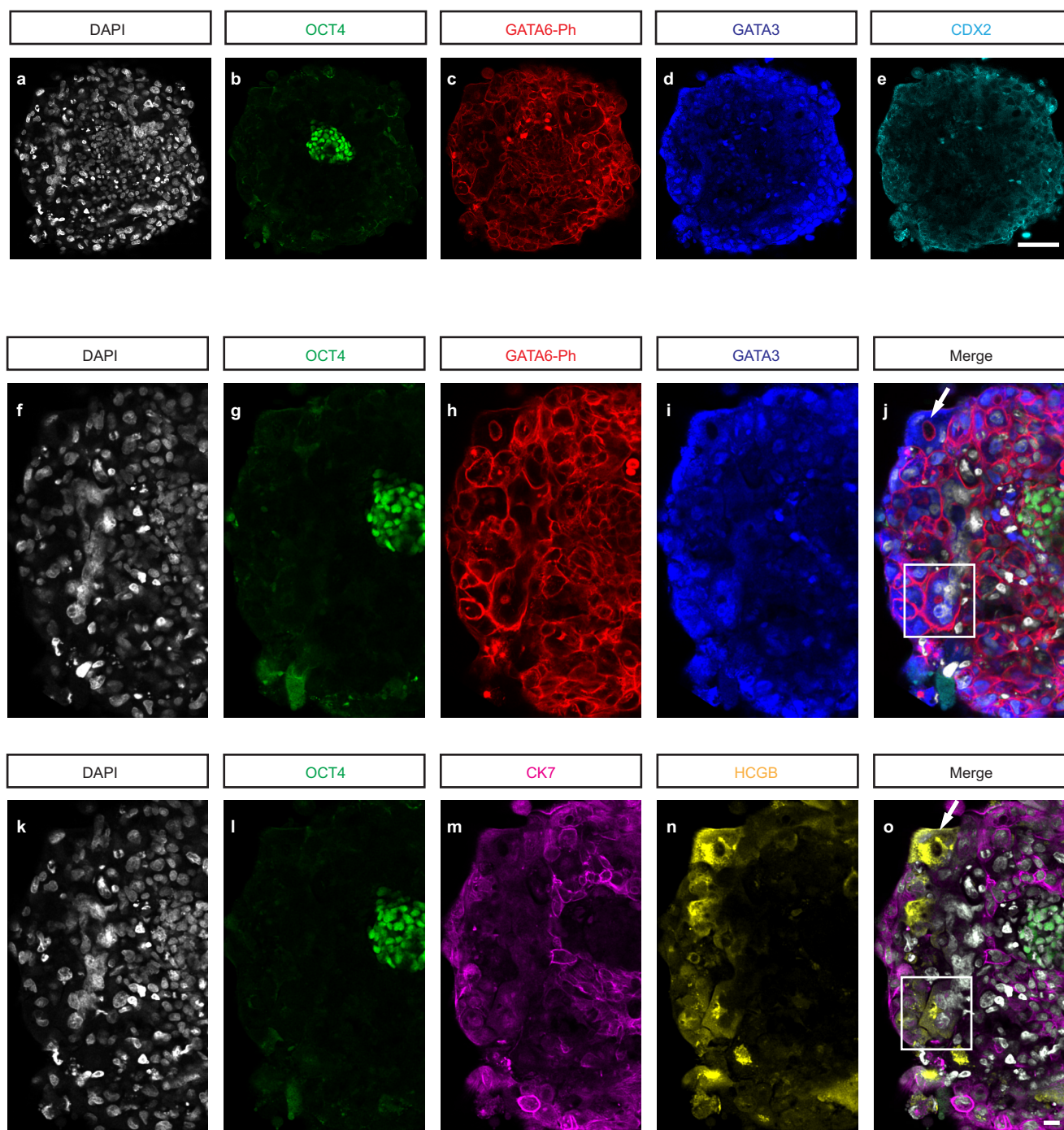


**Extended Data Figure 5 | Full z-series d.p.f. 10 amniotic and yolk sac cavities.** Series of confocal z-sections of the embryo from Fig. 3c stained for OCT4 (green), GATA6-phalloidin (red), and CDX2 (cyan), showing the amniotic and yolk sac cavities, with ysTE cells showing expression of varying levels of all three markers. The scale is the same as in Fig. 3d.





**Extended Data Figure 6 | Tilted three-dimensional reconstruction of a d.p.f. 12 embryo.** The embryo reconstruction is oriented at a 30° pitch, with individual channels shown in greyscale and then merged as OCT4 (green), GATA6–phalloidin (red), GATA3 (blue). Phalloidin staining of this embryo was so intense upon re-staining that the original underlying GATA6 nuclear stain is not visible in this data set. Scale bar, 50  $\mu$ m.



**Extended Data Figure 7 | Additional images from a d.p.f. 12 embryo.** **a–e**, Confocal z-section of the d.p.f. 12 embryo in Fig. 4c stained for DAPI (white), OCT4 (green), GATA6–phalloidin (GATA6-Ph, red), GATA3 (blue), and CDX2 (cyan). Scale bar, 100  $\mu$ m. **f–j**, Confocal sections of the same d.p.f. 12 embryo stained and imaged first (**f–j**) for DAPI (white), OCT4 (green), GATA6–phalloidin (red), and GATA3 (blue) and then (**k–o**) re-stained and re-imaged for DAPI (white), OCT4 (green),

CK7 (magenta), and HCGB (yellow). DAPI and OCT4 were used as landmarks to identify the same z-plane and the same cells between the two staining and imaging rounds. The arrow in **j** and **o** indicates an example of nascent lacuna, typical of ST cells at Carnegie stage 5c *in vivo*; the box indicates an example of multinucleated cells characteristic of ST lineage progression; zoom-ins of the area in the box are presented in Fig. 4r–v. Scale bar, 20  $\mu$ m.



# The evolution of cooperation within the gut microbiota

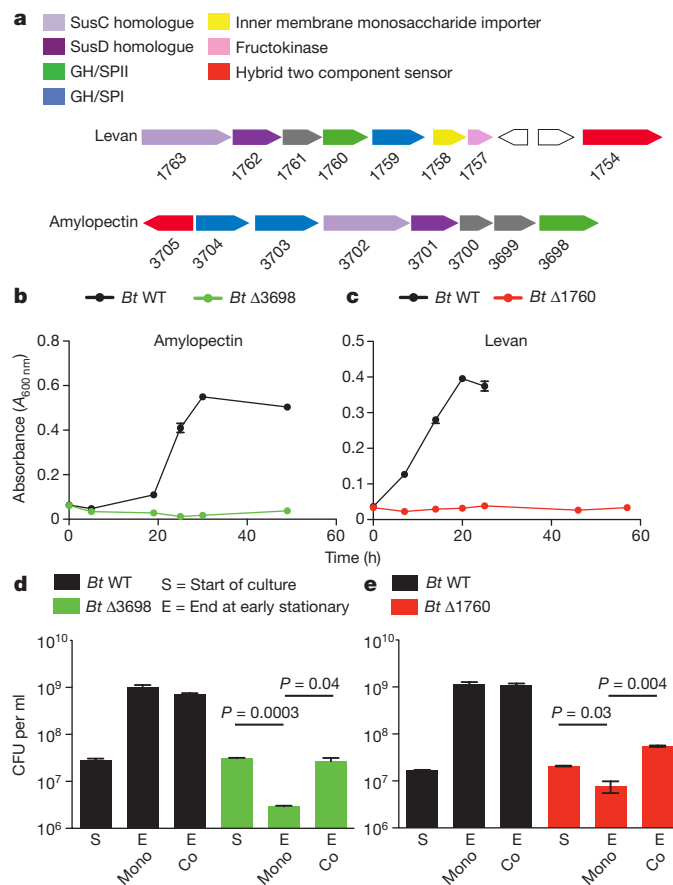
Seth Rakoff-Nahoum<sup>1,2</sup>, Kevin R. Foster<sup>3</sup> & Laurie E. Comstock<sup>2</sup>

Cooperative phenotypes are considered central to the functioning of microbial communities in many contexts, including communication via quorum sensing, biofilm formation, antibiotic resistance, and pathogenesis<sup>1–5</sup>. The human intestine houses a dense and diverse microbial community critical to health<sup>1,2,4–9</sup>, yet we know little about cooperation within this important ecosystem. Here we test experimentally for evolved cooperation within the Bacteroidales, the dominant Gram-negative bacteria of the human intestine. We show that during growth on certain dietary polysaccharides, the model member *Bacteroides thetaiotaomicron* exhibits only limited cooperation. Although this organism digests these polysaccharides extracellularly, mutants lacking this ability are outcompeted. In contrast, we discovered a dedicated cross-feeding enzyme system in the prominent gut symbiont *Bacteroides ovatus*, which digests polysaccharide at a cost to itself but at a benefit to another species. Using *in vitro* systems and gnotobiotic mouse colonization models, we find that extracellular digestion of inulin increases the fitness of *B. ovatus* owing to reciprocal benefits when it feeds other gut species such as *Bacteroides vulgatus*. This is a rare example of naturally-evolved cooperation between microbial species. Our study reveals both the complexity and importance of cooperative phenotypes within the mammalian intestinal microbiota.

A major challenge facing the study of host-associated microbiotas is to understand the ecological and evolutionary dynamics that shape these communities<sup>2,5,10–13</sup>. A key determinant of microbial dynamics is the balance of cooperation and competition both within and between species<sup>2,5,14,15</sup>. Here we test for the evolution of cooperation within the mammalian microbiota by focusing on the Bacteroidales, the most abundant order of Gram-negative bacteria of the human intestine with species that co-colonize the host at high densities of  $10^9$ – $10^{11}$  CFU per gram of faeces<sup>16,17</sup>. Members of this order break down polysaccharides outside of their cell using outer surface glycoside hydrolases<sup>6,18</sup>, some of which are secreted on outer membrane vesicles<sup>3,19</sup>. This suggests a significant potential for one cell to cooperatively feed other cells. As extracellular digestion is considered important for growth of Bacteroidales on polysaccharides<sup>1,6,8</sup>, we focused on this trait as a candidate for cooperative interactions within the gut microbiota and asked whether a bacterium that breaks down a polysaccharide extracellularly receives most, or all<sup>20</sup>, of the benefits of its efforts.

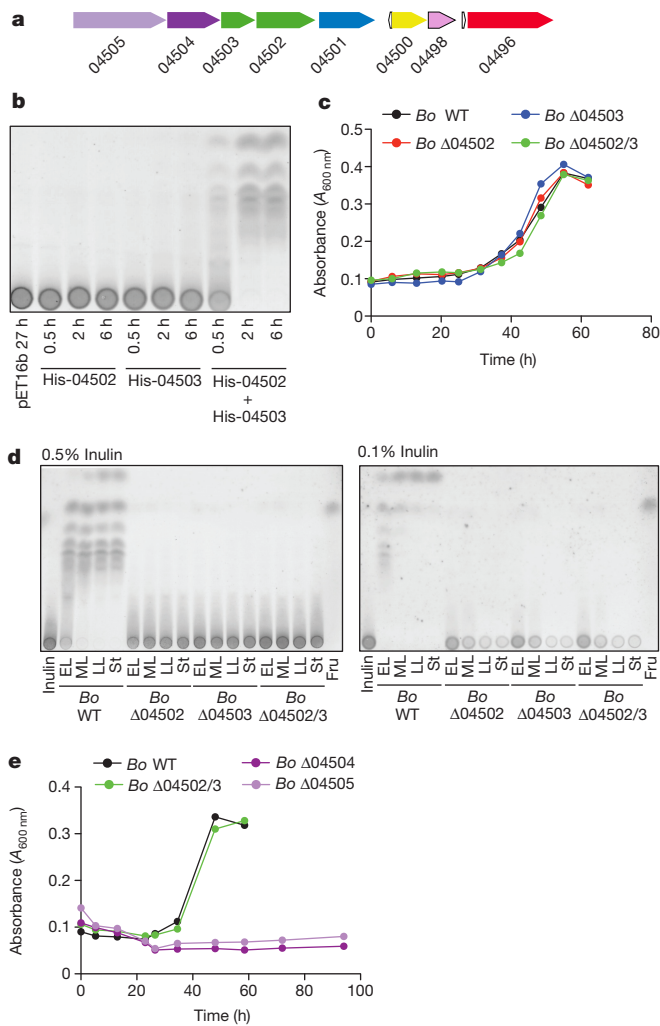
We first made isogenic mutants in genes responsible for the extracellular digestion of polysaccharides in the well-studied human gut strain, *Bacteroides thetaiotaomicron* VPI-5482 (refs 1, 8). Specifically, we deleted the genes encoding the outer surface glycoside hydrolases BT3698 of the amylopectin/starch utilization locus and BT1760 of the levan utilization locus (Fig. 1a), required for growth on amylopectin and levan, respectively (Fig. 1b, c, Extended Data Fig. 1a, b). Consistent with previous observations<sup>1,8</sup>, neither mutant grew with the specific polysaccharide in monoculture (Fig. 1b, c, Extended Data Fig. 1a, b). However, co-culture of  $\Delta$ BT3698 or  $\Delta$ BT1760 with wild type in amylopectin or levan

increased the fitness of the mutants (Fig. 1d, e, Extended Data Fig. 1c, d). This is consistent with cooperation via public good availability of amylopectin and levan breakdown products.



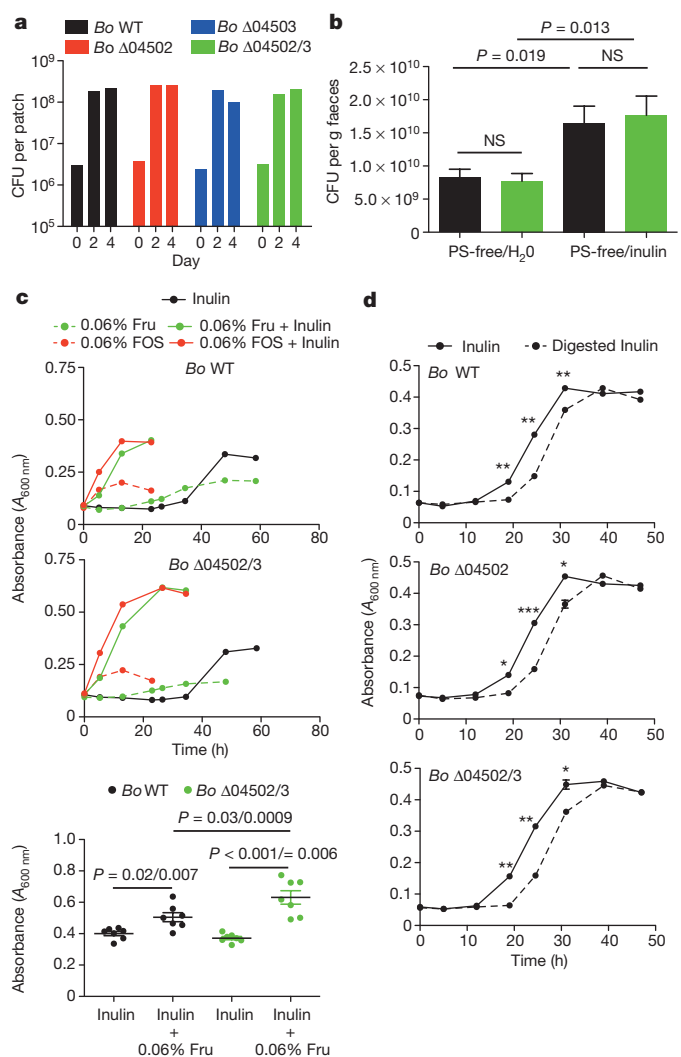
**Figure 1 | Direct and cooperative benefits of polysaccharide digestion by surface glycoside hydrolases (GH).** **a**, Polysaccharide utilization loci of *B. thetaiotaomicron* (*Bt*) for amylopectin and levan with the products or properties each gene encodes listed above and colour-coded. SPI or SPII, signal peptidase I or II cleavage site, respectively. **b**, **c**, Growth of *Bt* wild type (WT) and surface GH mutants in media with amylopectin ( $n = 2$ , cell culture biological replicates) (**b**) or levan ( $n = 2$ , cell culture biological replicates) (**c**). See Extended Data Fig. 1 for additional independent experiments. **d**, **e**, Growth of *Bt* WT and surface GH mutants in mono- and co-culture in media with amylopectin ( $n = 2$ , cell culture biological replicates) (**d**) or levan ( $n = 2$ , cell culture biological replicates) (**e**). See Extended Data Fig. 1 for additional independent experiments. In all panels error bars represent standard error;  $P$  values derived from two-tailed Student's  $t$ -test.

<sup>1</sup>Division of Infectious Diseases, Department of Medicine, Boston Children's Hospital and Harvard Medical School, 300 Longwood Avenue, Boston, Massachusetts 02115, USA. <sup>2</sup>Division of Infectious Diseases, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, 181 Longwood Avenue, Boston, Massachusetts 02115, USA. <sup>3</sup>Department of Zoology and Oxford Centre for Integrative Systems Biology, University of Oxford, Oxford OX1 3PS, UK.



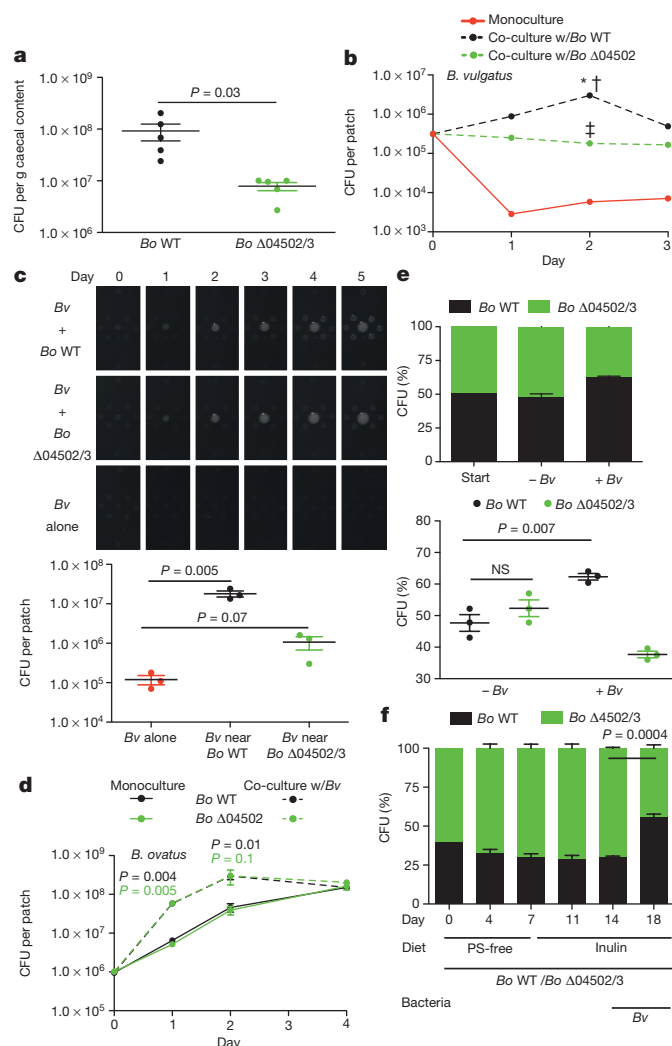
**Figure 2 | *B. ovatus* does not require surface digestion for utilization of inulin.** **a**, Predicted inulin utilization locus of *B. ovatus* (*Bo*). Gene designations shown are preceded by BACOVA\_. The colour coding of gene products is as in Fig. 1. **b**, Thin layer chromatography (TLC) analysis of inulin defined medium incubated for the indicated times with purified His-tagged 04502, 04503 or 04501 and 04503, or vector control. See Supplementary Information Fig. 1 for uncropped scanned images. **c**, Growth of *Bo* WT and mutants in inulin defined medium. **d**, TLC analysis of conditioned media during the growth of *Bo* WT and mutants in 0.5% and 0.1% inulin. EL, early log; ML, mid log; LL, late log; St, stationary. **e**, Growth of *Bo* WT,  $\Delta 04502/3$ ,  $\Delta 04504$  (*susD* orthologue) and  $\Delta 04505$  (*susC* orthologue) mutants in inulin defined medium. For **c**, **e**, each line represents  $n = 1$  sample per condition. See Extended Data Fig. 2 for additional independent experiments. In all panels error bars represent standard error;  $P$  values derived from two-tailed Student's  $t$ -test.

One of the key questions in evolutionary biology is how cooperative systems can be evolutionarily stable<sup>4,5,14,21,22</sup>. If certain cells invest in the production of an enzyme that helps others, what prevents these cells from being outcompeted by cells that consume the breakdown products without making the enzyme? In the *B. thetaiotaomicron* amylopectin and levan polysaccharide utilization systems, while receiving public goods from the wild type benefits the mutant cells (Fig. 1d, e, Extended Data Fig. 1e), they do not outcompete the wild type. Cells that make the enzyme receive more benefits than non-producing neighbouring cells. This observation suggests that a cell can utilize the majority of the polysaccharide that it breaks down, and that these private<sup>23</sup> benefits are central to the evolutionary stability of extracellular polysaccharide digestions in these systems.



**Figure 3 | Cost of inulin digestion by surface glycoside hydrolases.** **a**, Growth yield of *Bo* WT and mutants on inulin agarose plates. Each line represents  $n = 1$  sample per condition. See Extended Data Fig. 4 for additional independent experiments. **b**, Bacteria per gram of faeces of gnotobiotic mice seven days after monocolonization with *Bo* WT or  $\Delta 04502/3$  on a polysaccharide-free diet or supplemented with inulin ( $n = 5$  biological replicate mice per condition). **c**, Growth curves (upper and middle panels) and maximal  $A_{600\text{nm}}$  (bottom panel) of *Bo* WT or  $\Delta 04502/3$  in minimal media with carbon sources as indicated. Each line represents  $n = 1$  sample per condition. See Extended Data Fig. 4 for additional independent experiments. In the bottom panel,  $P$  values are displayed as paired (matched WT and  $\Delta 04502/3$  within same experiment) followed by unpaired (all experiments) two-tailed Student's  $t$ -test ( $n = 7$ , biological replicates). **d**, Growth of *Bo* WT or mutants in inulin or a stoichiometric equivalent amount of inulin breakdown products after digestion with purified 04502 and 04503 enzymes ( $n = 2$ , cell culture biological replicates); \* $P < 0.05$ ; \*\* $P < 0.01$ ; \*\*\* $P < 0.001$ ; see Extended Data Fig. 5 for additional independent experiments. In all panels error bars represent standard error;  $P$  values derived from two-tailed Student's  $t$ -test.

We extended our analysis to another prominent member of the human Bacteroidales known to extensively utilize polysaccharides, *Bacteroides ovatus* (*Bo*)<sup>24</sup>. During growth on inulin, a dietary fructan known for health-promoting effects<sup>25</sup>, *Bo* extracellularly digests and liberates considerable amounts of inulin breakdown products<sup>3</sup>. The predicted inulin utilization locus of the *Bo* type strain ATCC 8483 encodes two similar outer surface glycoside hydrolases, BACOVA\_04502 and BACOVA\_04503 (Fig. 2a), both of which are predicted to target the  $\beta 1,2$  inulin fructose polymer<sup>1,3</sup>. Both of these enzymes are required for



inulin breakdown (Fig. 2b). We therefore predicted that a single mutant of either BACOVA\_04502 or BACOVA\_04503 would be unable to grow on inulin. Surprisingly, neither of the single deletion mutants ( $\Delta$ 04502 or  $\Delta$ 04503) nor the double mutant ( $\Delta$ 04502/3) demonstrated impaired fitness with inulin as the sole carbohydrate source (Fig. 2c, Extended Data Fig. 2a), even at limiting concentrations (Extended Data Fig. 2a).

Given the importance of extracellular polysaccharide digestion for growth of *Bacteroides* on numerous polysaccharides<sup>1,6,8</sup> (Fig. 1, Extended Data Fig. 1), we predicted that *Bo* synthesizes other enzymes that breakdown inulin extracellularly, allowing the *Bo* mutants to grow on this polysaccharide. However, analysis of the growth media of  $\Delta$ 04502,  $\Delta$ 04503 and  $\Delta$ 04502/3 revealed no released inulin breakdown products, demonstrating that BACOVA\_04502 and BACOVA\_04503 are solely responsible for extracellular digestion of inulin (Fig. 2d, Extended Data Fig. 2b; see Extended Data Fig. 2c for complementation). BACOVA\_04504 and BACOVA\_04505 encode SusD and SusC orthologues respectively, which are predicted to function in inulin binding and the import of digestion products<sup>6,26</sup>. Deletion of BACOVA\_04504 or BACOVA\_04505 resulted in significant impairment of growth on inulin (Fig. 2e, Extended Data Fig. 3a; see Extended Data Fig. 3b for complementation). Growth of  $\Delta$ 04502,  $\Delta$ 04503 and  $\Delta$ 04502/3 in limiting concentrations of inulin revealed depletion of inulin (Fig. 2d, right panel; Extended Data Fig. 2b). Together, these data demonstrate that surface enzymes 04502 and 04503 are not needed for *Bo* to utilize inulin, and suggest that inulin is directly imported via 04504–04505 without prior extracellular digestion.

Why would *Bo* synthesize surface/secreted enzymes that potentially digest inulin outside of the cell if not necessary for its growth on the

**Figure 4 | Interspecies cooperation mediated by surface digestion of inulin is stabilized by reciprocal benefits.** **a**, Germ-free mice were monocolonized with *Bo* WT or  $\Delta$ 04502/3 and maintained on a polysaccharide-free diet supplemented with inulin and housed for 2 weeks. Mice were then gavaged with the caecal contents of conventionally raised mice. At day 5 post gavage, caecal contents were plated for enumeration of *Bo* ( $n = 5$  biological replicate mice per group). **b**, Enumeration of *Bv* in monoculture or co-culture with *Bo* WT or  $\Delta$ 04502 on defined inulin plates.  $*P = 0.001$  for number of *Bv* in *Bv/Bo* WT co-culture vs *Bv* alone;  $\dagger P = 0.001$  for number of *Bv* in *Bv/Bo* WT co-culture vs *Bv/Bo*  $\Delta$ 04502 co-culture;  $\ddagger P = 0.003$  for number of *Bv* in *Bv/Bo*  $\Delta$ 04502 co-culture vs *Bv* alone ( $n = 2$ , cell culture biological replicates). See Extended Data Figs 7 and 9 for additional independent experiments. **c**, (left) Photos of patches of *Bv* plated at varying distances around *Bo* WT or *Bo*  $\Delta$ 04502/3 or alone on inulin agarose plates. (right) Enumeration of *Bv* after five days of culture plated at the same distance to *Bo* WT or *Bo*  $\Delta$ 04502/3 or alone on inulin plates. See Extended Data Fig. 9 for additional independent experiments. **d**, Enumeration of *Bo* WT or  $\Delta$ 04502/3 in monoculture or co-culture with *Bv* on defined inulin plates.  $P$  values correlate to colour of line/genotype used for statistical analysis.  $P$  value indicates comparison of monoculture and co-culture for the given condition at the time-point indicated. The benefit *Bo* receives from *Bv* is most robust when starting with fewer *Bo*. Depicted are starting CFU of  $\sim 10^6$  *Bo*. ( $n = 3$ , cell culture biological replicates at day 1, 2;  $n = 2$ , biological replicates at day 4). See Extended Data Fig. 7 for additional independent experiments and Extended Data Fig. 7b for starting CFU of  $\sim 10^7$  *Bo*. **e**, Ratios of wild type and  $\Delta$ 04502/3 at the start and day 2 of culture when co-plated with or without *Bv* on inulin plates ( $n = 3$ , cell culture biological replicates). See Extended Data Fig. 9 for additional independent experiments. **f**, Ratios of WT and  $\Delta$ 04502/3 in the inoculum (day 0) and in faeces at various time points (days 4, 7, 11, 14, 18) post co-colonization of gnotobiotic mice ( $n = 3$  cell culture biological replicates) with *Bo* WT and  $\Delta$ 04502/3. Polysaccharide-free diet was supplemented with inulin at day 7. *Bv* was introduced at day 14. A Fisher exact test comparing the frequency of *Bo* WT and  $\Delta$ 04502/3 pre- (day 14) and post- (day 18) colonization with *Bv* was significant with a  $P$  value of 0.0001 for each individual mouse. *Bo* CFU in faeces were maximal after the switch to inulin diet and addition of *Bv* changed the abundance of *Bo* WT compared to  $\Delta$ 04502/3 but not total CFU of *Bo*. See Extended Data Fig. 9 for additional independent experiments. In all panels error bars represent standard error;  $P$  values displayed are derived from two-tailed Student's  $t$ -test.

polysaccharide? A key evolutionary explanation for the release of secreted products by microbes is that they feed clonemates in a manner that is beneficial at the level of the clonal group<sup>2,5,11,14,15</sup>. We hypothesized that the importance of extracellular digestion may be realized during spatially structured growth on plates where not all cells are in direct contact with the polysaccharide. However, mutant bacteria showed no significant differences in growth yield compared to wild type on defined inulin plates (Fig. 3a; Extended Data Fig. 4a). In addition, these enzymes were not required for optimal growth *in vivo* as wild type and  $\Delta$ 04502/3 showed equal colonization levels in gnotobiotic mice fed a polysaccharide-free diet with inulin added as the sole dietary polysaccharide (Fig. 3b). Therefore, we could find no evidence that inulin breakdown by 04502 and 04503 benefits *Bo* in three-dimensional growth or during monocolonization of the mammalian gut.

Bacteroidales polysaccharide utilization loci can be induced by monomers or oligomers of the cognate polysaccharide<sup>1,26</sup>. This raised the possibility that 04502 and 04503 may be important for optimal growth on inulin during induction. However, although addition of trace amounts of fructose monomers or oligosaccharides led to accelerated growth on inulin (Fig. 3c, Extended Data Fig. 4b), this did not require 04502/3 (Fig. 3c, Extended Data Fig. 4b). Rather than a benefit from the presence of the enzymes, we observed a cost based on reduced yield of the wild type compared to  $\Delta$ 04502/3 mutant during induced growth (Fig. 3c, Extended Data Fig. 4b) that was independent of a direct energetic cost of synthesis of 04502/3 (Figs 2c, e, 3c, d and Extended Data Figs 2a, 4b and 5a, b, c). We find that *Bo* grows better on undigested inulin than an equal concentration of inulin breakdown products (Fig. 3d, Extended Data Fig. 5a, b). In addition, the yield advantage to the



mutant does not occur under limiting inulin concentrations (Extended Data Fig. 5d,e). Furthermore, *Bo* preferentially consumes longer inulin digestion products over shorter oligomers and fructose (Fig. 2d, right panel, Extended Data Fig. 2b, right panel). Together, these data suggest that undigested inulin is the preferred substrate of *Bo*, and that extracellular digestion by  $\Delta 04502/3$  under certain conditions is costly for fitness.

We found no evidence that extracellular digestion of inulin by *Bo* evolved for cooperation with clonemates. Therefore, we speculated that this trait might have evolved for cooperation with other species in the gut. We first sought evidence of cooperation in the setting of a natural gut ecosystem. Germ-free mice were fed inulin and colonized with either *Bo* wild type or  $\Delta 04502/3$  followed by the introduction of the caecal microbiota of conventionally raised mice. *Bo* wild type and  $\Delta 04502/3$  equally colonized mice before the introduction of the microbiota (Extended Data Fig. 6a), but *Bo* wild type received a significant fitness benefit compared to  $\Delta 04502/3$  in the context of a complex microbiota (Fig. 4a). These data suggested that, although not required for *Bo* to utilize inulin,  $\Delta 04502$  and  $\Delta 04503$  provide a benefit to *Bo* only realized in a community setting. The conditions for the evolution of cooperation between species are much more restrictive than those within a clone. In particular, theory predicts that costly interspecies cooperation will only be stabilized if there are reciprocal feedback benefits, such as a plant providing nectar for an insect that pollinates it<sup>21,22</sup>. From this experiment, we identified two dominant mouse microbiota Bacteroidales strains that thrived on *Bo*-derived inulin breakdown products (Extended Data Fig. 6b), with delayed growth on inulin (Extended Data Fig. 6b). These data suggest that cross-feeding Bacteroidales members may provide reciprocal benefits to wild-type *Bo* in the mammalian gut.

To test experimentally for reciprocity and benefits of inulin digestion between species, we used an inulin co-culture system with *Bo* wild type or  $\Delta 04502$  and *Bacteroides vulgatus* ATCC 8482 (*Bv*), which is commonly found together with *Bo* at high densities in humans<sup>16,17</sup> and thrives on inulin breakdown products<sup>3</sup> but cannot use inulin<sup>1,3</sup>. Co-culture and proximate plating with *Bo* wild type increased the fitness of *Bv* compared to that with  $\Delta 04502$  (Fig. 4b, c and Extended Data Figs 7a, 9a); however, *Bv* is able to persist better with  $\Delta 04502$  than when alone (Fig. 4b, Extended Data Fig. 7a), owing to a small (<2,000 Da) secreted molecule(s) that contributes to the survival of *Bv* (Extended Data Fig. 8a, b). This  $\Delta 04502/3$ -independent survival is not mediated by a universal factor made by *Bacteroides* during growth on inulin nor *Bo*-derived short-chain fatty acids (Extended Data Fig. 8a, c, d). Thus, there are multiple mechanisms by which *Bo* helps *Bv* (Fig. 4b, Extended Data Fig. 7a), the greatest being cross-feeding mediated by  $\Delta 04502/3$ .

We next addressed the question of whether *Bo* receives reciprocal benefits from *Bv*. Co-culture of *Bv* with *Bo* on plates increased the fitness of *Bo* wild type and *Bo*  $\Delta 04502$  (Fig. 4d, Extended Data Fig. 7b), but did not increase the fitness of *B. fragilis* (Extended Data Fig. 7c). If inulin breakdown can be costly, and *Bo* receives benefits from *Bv* irrespective of whether inulin is broken down and fed to *Bv*, natural selection is expected to favour the loss of the genes encoding the secreted inulin glycoside hydrolases<sup>22</sup>. However, our pairwise experiments would not reveal the possibility that *Bo* wild type receives more reciprocal benefits from *Bv* when in direct competition with the non-cross-feeding mutant<sup>21,27</sup>. Therefore, we co- and tri-cultured these strains on plates and compared the yields of the two *Bo* strains (wild type and  $\Delta 04502/3$ ) with or without *Bv*. Addition of *Bv* leads to an increased proportion of *Bo* wild type compared to the  $\Delta 04502/3$  mutant (Fig. 4e, Extended Data Fig. 9b). We extended these studies to a gnotobiotic mouse colonization model. *Bo* wild type had no advantage in direct competition with  $\Delta 04502/3$  on a polysaccharide-free diet or when inulin is added (Fig. 4f). However, introduction of *Bv* increased the fitness of the *Bo* wild type relative to the mutant (Fig. 4f, Extended Data Fig. 9c, d). Together, these data suggest that extracellular breakdown of inulin

increases the fitness of *Bo* via reciprocal benefits from another species. These findings are consistent with the evolution of cooperation between species within the gut microbiota.

We find evidence of distinct forms of cooperativity within the Bacteroidales of the human intestinal microbiota (Extended Data Fig. 10). For *B. thetaiotaomicron*, amylopectin and levan digestion provide mostly private benefits and modest social benefits to other cells. By contrast, *Bo* releases large amounts of inulin digestion products via a pair of dedicated cross-feeding secreted enzymes unnecessary for its use of inulin. These enzymes allow for cooperation with cross-fed species, which provide benefits in return. Potential mechanisms by which *Bv* may provide return benefits to *Bo* include detoxification of inhibitory substances, or production of a depleted or growth promoting factor, the latter supported by early growth benefits to *Bo* via *Bv* secreted factors (Extended Data Fig. 9e).

Understanding whether microbial communities are formally cooperative is central to predicting their evolutionary and ecological stability. Cooperative systems can be productive, but are prone to instabilities on both ecological and evolutionary timescales that can undermine them<sup>5,14,15,21,22</sup>. The ability of one species to utilize the waste product of another is prevalent, but waste production alone does not signify cooperative evolution. As opposed to waste product utilization or exploitive interactions<sup>28–30</sup>, there are few well-documented cases of evolved cooperation between microbial species where one species evolves to help another<sup>14</sup>. We have found evidence of strong eco-evolutionary interactions within the microbiota that are likely to be central to the functioning of these complex communities.

**Online Content** Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

**Received 4 May 2015; accepted 24 February 2016.**

**Published online 25 April 2016.**

1. Sonnenburg, E. D. *et al.* Specificity of polysaccharide use in intestinal bacteroides species determines diet-induced microbiota alterations. *Cell* **141**, 1241–1252 (2010).
2. West, S. A., Griffin, A. S., Gardner, A. & Diggle, S. P. Social evolution theory for microorganisms. *Nature Rev. Microbiol.* **4**, 597–607 (2006).
3. Rakoff-Nahoum, S., Coyne, M. J. & Comstock, L. E. An ecological network of polysaccharide utilization among human intestinal symbionts. *Curr. Biol.* **24**, 40–49 (2014).
4. Drescher, K., Nadell, C. D., Stone, H. A., Wingreen, N. S. & Bassler, B. L. Solutions to the public goods dilemma in bacterial biofilms. *Curr. Biol.* **24**, 50–55 (2014).
5. Frank, S. A. A general model of the public goods dilemma. *J. Evol. Biol.* **23**, 1245–1250 (2010).
6. Koropatkin, N. M., Cameron, E. A. & Martens, E. C. How glycan metabolism shapes the human gut microbiota. *Nature Rev. Microbiol.* **10**, 323–335 (2012).
7. Subramanian, S. *et al.* Cultivating healthy growth and nutrition through the gut microbiota. *Cell* **161**, 36–48 (2015).
8. Shipman, J. A., Cho, K. H., Siegel, H. A. & Salyers, A. A. Physiological characterization of SusG, an outer membrane protein essential for starch utilization by *Bacteroides thetaiotaomicron*. *J. Bacteriol.* **181**, 7206–7211 (1999).
9. Littman, D. R. & Pamer, E. G. Role of the commensal microbiota in normal and pathogenic host immune responses. *Cell Host Microbe* **10**, 311–323 (2011).
10. Waldor, M. K. *et al.* Where next for microbiome research? *PLoS Biol.* **13**, e1002050 (2015).
11. Koschwanez, J. H., Foster, K. R. & Murray, A. W. Sucrose utilization in budding yeast as a model for the origin of undifferentiated multicellularity. *PLoS Biol.* **9**, e1001122 (2011).
12. Costello, E. K., Stagaman, K., Dethlefsen, L., Bohannan, B. J. M. & Relman, D. A. The application of ecological theory toward an understanding of the human microbiome. *Science* **336**, 1255–1262 (2012).
13. Estrela, S., Whiteley, M. & Brown, S. P. The demographic determinants of human microbiome health. *Trends Microbiol.* **23**, 134–141 (2015).
14. Mitri, S. & Foster, K. R. The genotypic view of social interactions in microbial communities. *Annu. Rev. Genet.* **47**, 247–273 (2013).
15. Oliveira, N. M., Niehus, R. & Foster, K. R. Evolutionary limits to cooperation in microbial communities. *Proc. Natl Acad. Sci. USA* **111**, 17941–17946 (2014).
16. Faith, J. J. *et al.* The long-term stability of the human gut microbiota. *Science* **341**, 1237439 (2013).

17. Zitomersky, N. L., Coyne, M. J. & Comstock, L. E. Longitudinal analysis of the prevalence, maintenance, and IgA response to species of the order Bacteroidales in the human gut. *Infect. Immun.* **79**, 2012–2020 (2011).
18. Flint, H. J., Bayer, E. A., Rincon, M. T., Lamed, R. & White, B. A. Polysaccharide utilization by gut bacteria: potential for new insights from genomic analysis. *Nature Rev. Microbiol.* **6**, 121–131 (2008).
19. Elhenawy, W., Debelyy, M. O. & Feldman, M. F. Preferential packing of acidic glycosidases and proteases into Bacteroides outer membrane vesicles. *mBio* **5**, e00909–14 (2014).
20. Cuskin, F. *et al.* Human gut Bacteroidetes can utilize yeast mannan through a selfish mechanism. *Nature* **517**, 165–169 (2015).
21. Sachs, J. L., Mueller, U. G., Wilcox, T. P. & Bull, J. J. The evolution of cooperation. *Q. Rev. Biol.* **79**, 135–160 (2004).
22. Foster, K. R. & Wenseleers, T. A general model for the evolution of mutualisms. *J. Evol. Biol.* **19**, 1283–1293 (2006).
23. Gore, J., Youk, H. & van Oudenaarden, A. Snowdrift game dynamics and facultative cheating in yeast. *Nature* **459**, 253–256 (2009).
24. Martens, E. C. *et al.* Recognition and degradation of plant cell wall polysaccharides by two human gut symbionts. *PLoS Biol.* **9**, e1001221 (2011).
25. Orel, R. & Kamhi Trop, T. Intestinal microbiota, probiotics and prebiotics in inflammatory bowel disease. *World J. Gastroenterol.* **20**, 11505–11524 (2014).
26. Cameron, E. A. *et al.* Multifunctional nutrient-binding proteins adapt human symbiotic bacteria for glycan competition in the gut by separately promoting enhanced sensing and catalysis. *mBio* **5**, e01441–14 (2014).
27. Momeni, B., Waite, A. J. & Shou, W. Spatial self-organization favors heterotypic cooperation over cheating. *Elife* **2**, e00960 (2013).
28. Ng, K. M. *et al.* Microbiota-liberated host sugars facilitate post-antibiotic expansion of enteric pathogens. *Nature* **502**, 96–99 (2013).
29. Degnan, P. H., Barry, N. A., Mok, K. C., Taga, M. E. & Goodman, A. L. Human gut microbes use multiple transporters to distinguish vitamin B<sub>12</sub> analogs and compete in the gut. *Cell Host Microbe* **15**, 47–57 (2014).
30. Fischbach, M. A. & Sonnenburg, J. L. Eating for two: how metabolism establishes interspecies interactions in the gut. *Cell Host Microbe* **10**, 336–347 (2011).

**Supplementary Information** is available in the online version of the paper.

**Acknowledgements** We thank C. B. Ogbunugafor, J. Ordovas-Montanes, U. von Andrian and M. Waldor for suggestions. M. Delaney for SCFA analysis, V. Yeliseyev for assistance with gnotobiotics. Inulin and FOS were provided by Beneo-Orafti. Mice were provided by the HDDC, NIH Grant P30 DK34845. S.R.-N. is supported by the PIDS-St Jude Research Hospital Fellowship Program in Basic Research, a K12 Child Health Research Center grant through Boston Children's Hospital and a Pilot Feasibility Award funded by HDDC P30 DK034854. K.R.F. is supported by European Research Council Grant 242670. This work was supported by Public Health Service grant R01AI081843 (to L.E.C.) from the NIH/NIAID.

**Author Contributions** S.R.-N. performed mutant construction bacterial cultures, gnotobiotic experiments, protein purification and TLC, L.E.C. assisted with mutant construction. S.R.-N. and L.E.C. analysed the data. S.R.-N., K.R.F. and L.E.C. designed the study and wrote the paper.

**Author Information** Reprints and permissions information is available at [www.nature.com/reprints](http://www.nature.com/reprints). The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to S.R.-N. ([seth.rakoff-nahoum@childrens.harvard.edu](mailto:seth.rakoff-nahoum@childrens.harvard.edu)).

## METHODS

**Bacterial strains and media.** Bacteroidales type strains used in this study are *Bo* ATCC 8483, *B. thetaiotaomicron* VPI 5482 *Bv* ATCC 8482, and *B. fragilis* NCTC 9343. Bacteria were grown in media formulation as previously described<sup>3</sup>.

**Bacterial culture.** For growth in defined media, bacteria were inoculated from brain heart infusion plates containing haemin and vitamin K (BHIS) plates into basal medium (BS), cultured overnight to stationary phase, then diluted 1:10 in fresh BS and grown to mid log. At mid log, bacteria were pelleted by centrifugation and washed with sterile phosphate buffered saline (PBS) and then inoculated in defined media. Carbohydrates used to supplement defined media include fructose (F2543, Sigma), fructose oligosaccharides (FOS; OrafitP95, Beneo-Orafti group), levan (L8647, Sigma), amylopectin (10120, Sigma), and inulin (OrafitHP, Beneo-Orafti group). Levan and amylopectin were autoclaved as 1% w/v in H<sub>2</sub>O and dialysed using 3.5 kDa MW membranes (Slide-A-Lyzer Dialysis Cassettes, ThermoScientific). Short chain fatty acids acetate, propionate and succinate were purchased from Sigma. Stock solutions of 2 mM were pH neutralized to pH 7.2–7.3 with 10 N NaOH. All cultures were grown at 37°C under anaerobic conditions. Bacterial growth was quantified by absorbance ( $A_{600\text{ nm}}$ ) using 200 µl of bacterial culture in 96-well flat-bottom microtitre plates using a Powerwave spectrophotometer (Biotek). Murine gut Bacteroidales from the caecal preparations used in for the colonization experiments were grown on BHIS plates. Resulting colonies were tested for growth in inulin minimal medium or inulin breakdown products from the conditioned media of *Bo* grown in inulin, containing inulin breakdown products as previously described<sup>3</sup>.

**Creation of *Bacteroides* mutants.** Deletion mutants were created whereby the genes encoding BT3698 or BT1760 in *Bacteroides thetaiotaomicron* VPI 5482, BACOVA\_04502 in *Bo* ATCC 8483 BACOVA\_04503, BACOVA\_04502/3, BACOVA\_04504, or BACOVA\_04505 in *Bo* ATCC 8483 were removed. DNA segments upstream and downstream of the region to be deleted were PCR amplified using the primers outlined in Supplementary Table 1. PCR products were digested with BamHI, EcoRI and/or MluI engineered into the primers (Supplementary Information Table 1) and cloned by three-way ligation into the appropriate site of pNJR6 (ref. 31). The resulting plasmids were conjugally transferred into the *Bacteroides* strain as indicated using helper plasmid R751 and cointegrates were selected by erythromycin resistance. Cross outs were screened by PCR for the mutant genotype.

**Cloning of PUL genes for expression in deletion mutants.** BACOVA\_04502, BACOVA\_04503, BACOVA\_04502/3, BACOVA\_04504, BACOVA\_04505 or BACOVA\_04504/5 genes were PCR amplified using the primers listed in Supplementary Table 1. The PCR products were digested and ligated into the BamHI or KpnI site of the *Bacteroides* expression vector pFD340<sup>32</sup>. Plasmids containing the correct orientation of the insert in relation to the vector-borne promoter were introduced into mutant *Bacteroides* strains by mobilization from *E. coli* using helper plasmid RK231.

**Mono-, co- and tri- culture experiments.** For bacterial mono and co-culture experiments in defined liquid media, bacteria were grown as indicated for monoculture before addition to the defined media. Sterile magnetic stir bars were added to culture tubes within a rack placed on a stir plate within the anaerobic chamber. For conditioned media experiments, *Bo* WT, *Bo* Δ04502, *Bo* Δ04502/3, *Bv*, or *Bf* were grown to early log in inulin defined media or 0.125% fructose defined media (for Extended Data Fig. 9e), conditioned media were collected, filter sterilized, and incubated at 37°C for 72 h. Conditioned media was replenished with defined media without additional carbohydrate and used for cultivation of *Bv*. *Bo* Δ04502/3 conditioned media was dialysed in defined media without carbohydrate using 2 kDa MW membranes (Slide-A-Lyzer Dialysis Cassettes, ThermoScientific). For monoculture of *Bo* WT and mutants on solid agarose, 4 µl of the indicated concentration of bacteria were dotted onto minimal inulin agarose plates. At the indicated time points, the bacteria were cut out, diluted and plated onto BHIS for CFU enumeration. For co-plating experiments, 10<sup>6</sup> *Bo* (WT or mutant) or *Bf* were co-plated with 10<sup>5</sup> *Bv* or a control volume of PBS. Four µl were then dotted on to defined inulin agarose plates. At the indicated time points, the dotted patches were cut from the agarose plates and resuspended in PBS, diluted and plated to BHIS for enumeration. Quantification and differentiation of wild type and isogenic glycoside hydrolase or polysaccharide lyase mutant was performed by plating dilutions of mixed liquid culture or the cut-out patch on agarose plates onto BHIS, followed by picking ~100 colonies and determining wild type or mutant genotype by PCR using primers listed in Supplementary Table 1. For genotypic screening of *B. thetaiotaomicron* wild type and *B. thetaiotaomicron* Δ3698, two sets of primers were used (Supplementary Table 1). For *Bo*/*Bv* co-culture in Fig. 4b, c and Extended Data Fig. 7, *Bo* Δ03533 (WT) and *Bo* Δ03533 Δ04502 (Δ04502) arginine auxotrophic mutants were used in co-culture with *Bv* on minimal inulin agarose plates supplemented with 50 µg ml<sup>-1</sup> of arginine (Sigma) which does not impair or limit

growth as compared to wild type<sup>3</sup>. Colonies on BHIS plates were replica plated onto defined glucose defined plates, which support the growth of *Bv* but not *Bo* Δ03533 or *Bo* Δ03533 Δ04502.

**Thin-layer chromatography.** Thin-layer chromatography (TLC) was employed to specifically detect carbohydrates as previously described<sup>3</sup>. Standards for TLC included glucose (G7528, Sigma), fructose (F2543, Sigma), fructose oligosaccharides (FOS; OrafitP95, Beneo-Orafti group) and inulin (OrafitHP, Beneo-Orafti group). See Supplementary Information Fig. 1 for uncropped TLCs.

**Gas chromatographic analysis of culture media.** Chromatographic analysis was carried out using a Shimadzu GC14-A system with a flame ionization detector (FID) (Shimadzu Corp, Kyoto, Japan). A volatile acid mix containing 10 mM of acetic, propionic, isobutyric, butyric, isovaleric, valeric, isocaproic, caproic, and heptanoic acids was used (Matreya, Pleasant Gap PA). A non-volatile acid mix containing 10 mM of pyruvic and lactic acid and 5 mM of oxaloacetic, oxalic, methyl malonic, malonic, fumaric, and succinic acid was used (Matreya, Pleasant Gap PA).

**Cloning, purification, and enzymatic analysis of BACOVA\_04502-3.** To obtain purified BACOVA\_04502 and BACOVA\_04503 proteins, these genes were cloned individually into the BamHI site of pET16b (Novagen) using the primers listed in Supplementary Table 1. The constructs were designed so that the His-tag encoded by pET16b replaced the SpII signal sequence of these proteins allowing for their solubility. The recombinant plasmids were transformed into *E. coli* BL21 (DE3), grown to an  $A_{600\text{ nm}}$  of 0.6–0.7, and expression of the recombinant gene was induced by the addition of 0.4 mM IPTG for an additional 4 h. The His-tagged proteins were isolated essentially as described<sup>33</sup> using Dynabeads TALON paramagnetic beads. For enzymatic analysis (Fig. 2a) the proteins were added to inulin media in their magnetic bead-bound form. This allows for easy removal of the enzymes following digestion. As a control, the beads resulting from the same procedure performed with *E. coli* BL21 (DE3) containing only the vector (pET16b) were used. For *Bo* growth assays (Fig. 3d), 50 µl of beads (25 µl containing His-04502 and 25 µl containing His-04503) or equivalent volume of Dynabead buffer (for undigested inulin) were added to inulin defined medium. After 24 h at 37°C, the beads containing the enzyme were removed with a magnet, and the media (digested or undigested inulin) was used to culture *Bo* WT and Δ04502.

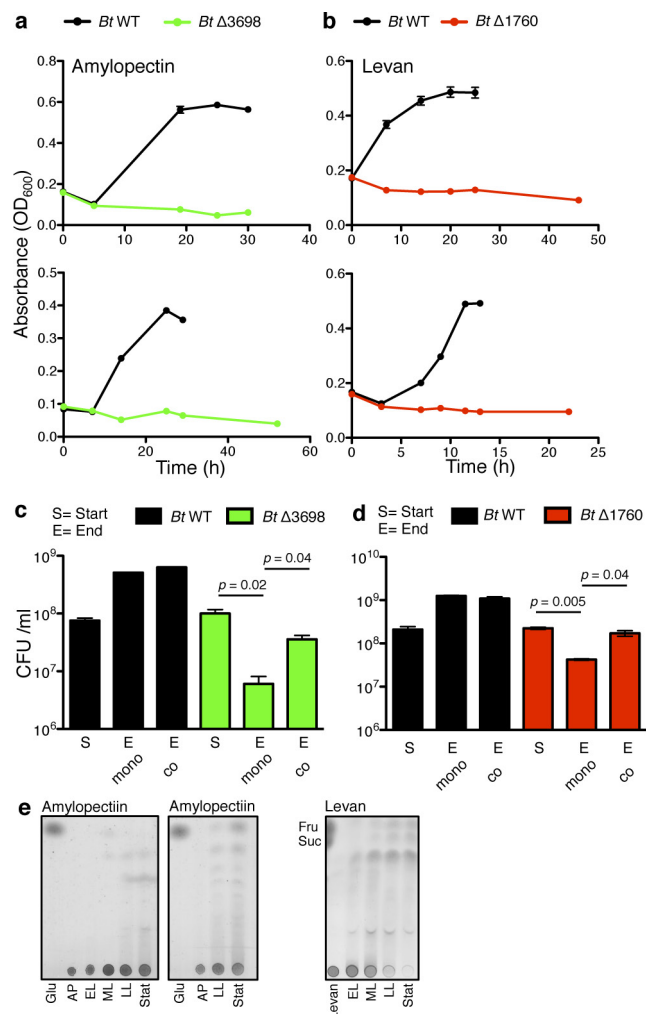
**Gnotobiotic mouse experiments.** All animal experiments were approved by the Harvard Medical School IACUC. Swiss Webster germ-free male mice (6–10 weeks old) were purchased from the Harvard Digestive Diseases Gnotobiotic Core facility. Littermates were randomly allocated for different gnotobiotic experimental arms. Experiments were conducted in either sterile Optima cages (Figs 3b, 4a, Extended Data Fig. 9d) or gnotobiotic isolators (Fig. 4f). When appropriate, animals were numbered by sterile tail markers for longitudinal analysis and were not blinded. Number of animals used for experimentation was determined by precedence for gnotobiotic studies. Longitudinal analysis in the experiment of Fig. 4f of the ratios of wild type to mutant *Bo* under changing environmental conditions allowed for internally-controlled, within-individual comparisons. Mice were placed on polysaccharide-free special chow (65% w/v glucose, protein-free, supplemented with all essential amino acids except arginine; BioServe). Arginine at 50 µg ml<sup>-1</sup> was supplemented in all drinking water. As indicated, mice were given 1% inulin (w/v) in sterile drinking water. Mice were inoculated with the indicated bacteria by applying ~10<sup>8</sup> live bacteria (grown to mid log) onto mouse fur. Dilutions of faeces at various time points following inoculation were plated to BHIS plates and genotyped as above (Fig. 4f). Predetermined exclusion criteria for gnotobiotic experiments contamination as determined by either the presence of colonies with distinct morphology on anaerobic plates than *Bo* or *Bv* or colonies present at >10<sup>2</sup> CFU ml<sup>-1</sup> (limit of detection) under aerobic conditions.

For gavage of *Bo* WT or Δ04502/3 monocolonized mice with the caecal content of conventionalized raised mice, two ~8-week-old male Swiss Webster mice purchased from Taconic and housed in a specific pathogen-free facility were euthanized under sterile conditions. Intestine was excised and care was made to leave caecum intact. Within 2 min of excision, the caeca were transferred into an anaerobic chamber where the caecal contents were pooled and diluted with ~10 ml of pre-reduced phosphate buffered solution supplemented with 0.1% cysteine. Contents were vigorously vortexed and immediately divided equally into two conical flasks (one for each group of mouse) to ensure equality of suspension between both conical flasks. Tightly sealed conical flasks were transferred to the mouse facility at which point 200 µl were gavaged to each mouse housed in sterile Optima cages. Caecal contents were plated to LKV (Remel) plates for enrichment of Bacteroides. Different colony morphologies were assigned species by 16S PCR as previously described<sup>17</sup>. For each group of mice, mice were housed in cages of 2 and 3 mice per cage. At time of killing, caecal contents were collected and plated to defined inulin plates by which *Bo* was enumerated based on distinct colony morphology of *Bo* on inulin agarose plates that was not present in conventionally raised caecal population.

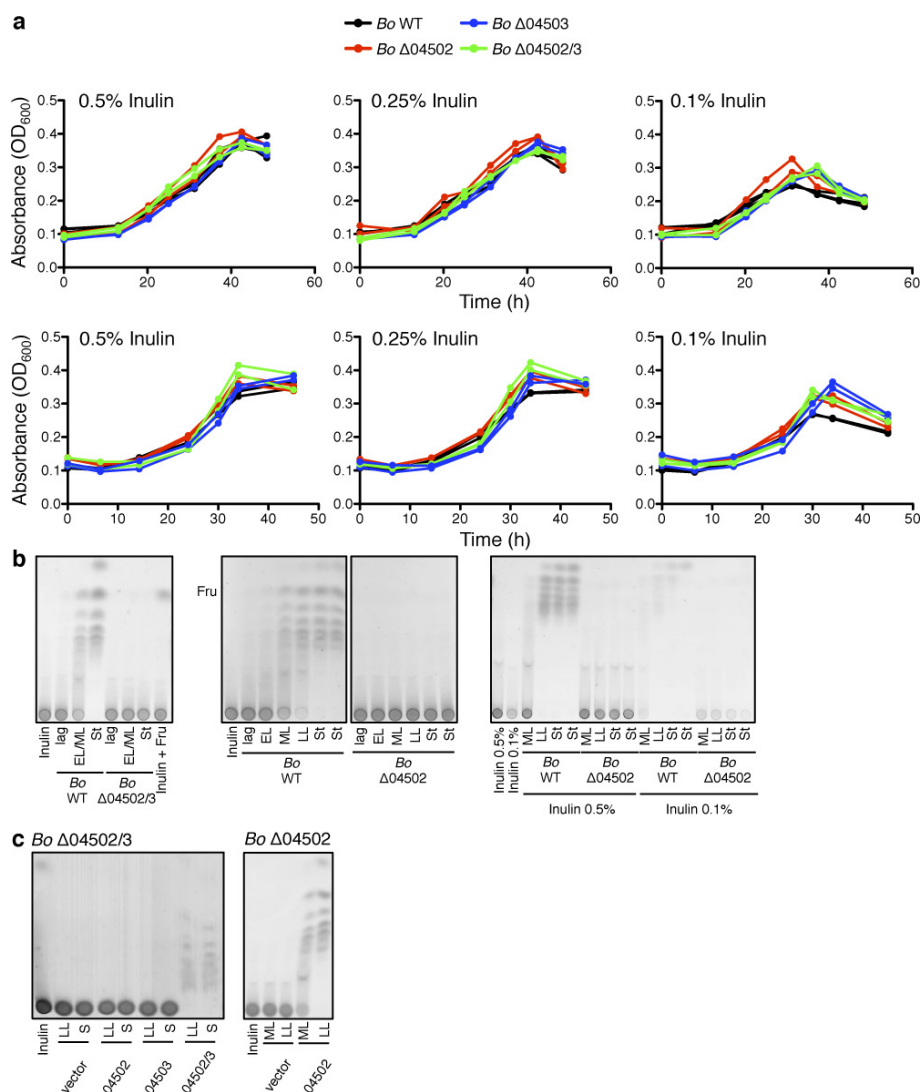


**Statistical analysis.** The experiments were not randomized, and the investigators were not blinded to allocation during experiments and outcome assessment. No statistical methods were used to predetermine sample size. Replicate experiments are shown in Extended Data. All *P* values are derived from Student's *t*-test except as indicated in Fig. 4f and Extended Data Fig. 9c where Fisher's exact test was performed. Statistical significance of variance reported as indicated per experiment in figure legends. All centre values are mean. Error bars are standard error of the mean.

31. Stevens, A. M., Shoemaker, N. B. & Salyers, A. A. The region of a *Bacteroides* conjugal chromosomal tetracycline resistance element which is responsible for production of plasmidlike forms from unlinked chromosomal DNA might also be involved in transfer of the element. *J. Bacteriol.* **172**, 4271–4279 (1990).
32. Smith, C. J., Rogers, M. B. & McKee, M. L. Heterologous gene expression in *Bacteroides fragilis*. *Plasmid* **27**, 141–154 (1992).
33. Coyne, M. J., Fletcher, C. M., Reinap, B. & Comstock, L. E. UDP-glucuronic acid decarboxylases of *Bacteroides fragilis* and their prevalence in bacteria. *J. Bacteriol.* **193**, 5252–5259 (2011).



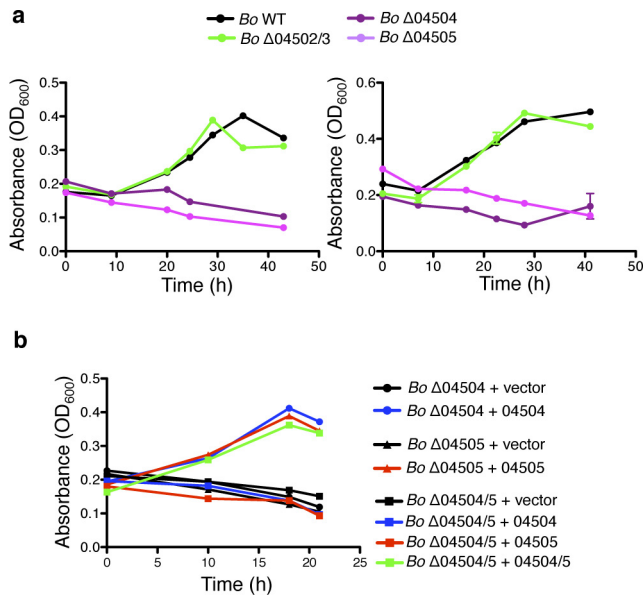
**Extended Data Figure 1 | Limited cooperation during polysaccharide utilization by *B. thetaiotaomicron*.** **a–d**, Independent experiments for Fig. 1b–e. **a**, **b**, Upper panels  $n = 2$  biological replicates, lower panels  $n = 1$ ; **c**, **d**,  $n = 2$  biological replicates. **e**, TLC analysis of conditioned media from *B. thetaiotaomicron* grown in amylopectin (left panel) or levan (right panel) minimal media. EL, early log; ML, late log; LL, late log; Stat, stationary phase; Glu, glucose; Fru, fructose; Suc, sucrose. See Supplementary Information Fig. 1 for uncropped scanned images. In all panels, error bars represent standard error;  $P$  values derived from two-tailed Student's  $t$ -test.



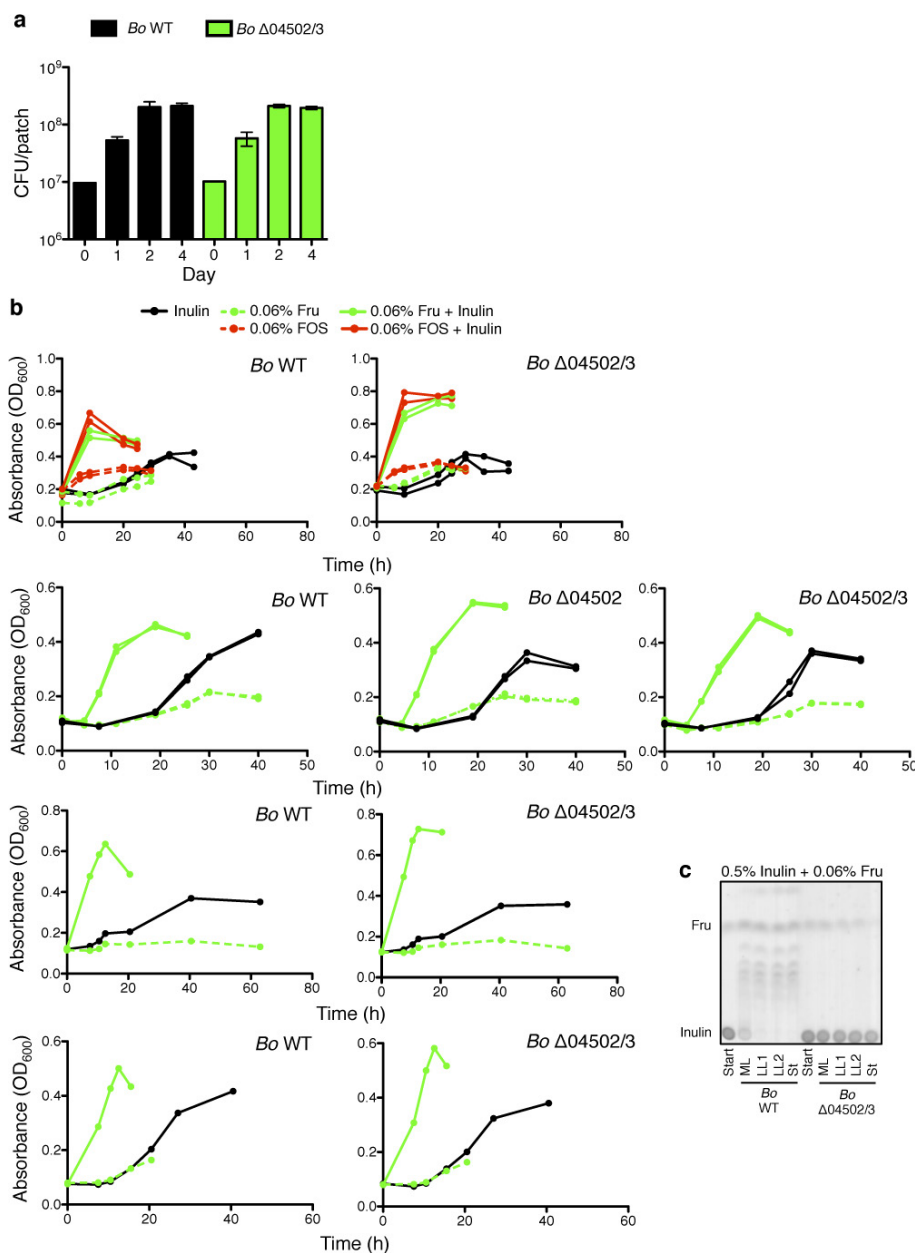
**Extended Data Figure 2 | *Bo* 04502 and 04503 mutants grow equivalently to wild type on limiting concentrations of inulin and do not require surface digestion for utilization of inulin.** **a**, Growth of *Bo* WT,  $\Delta$ 04502,  $\Delta$ 04503 and  $\Delta$ 04502/3 mutants in varying concentrations of inulin as indicated. Biological replicates of each condition are plotted as individual lines ( $n = 2$  cell culture biological replicates). Upper and lower panels are independent experiments. **b**, Independent experiments for Fig. 2d. EL, early log; ML, mid log; LL, late log; St, stationary phase.

See Supplementary Information Fig. 1 for uncropped scanned images. **c**, Complementation of *Bo*  $\Delta$ 04502 and  $\Delta$ 04502/3 mutants with the respective genes *in trans*. TLC analysis of conditioned media from *Bo*  $\Delta$ 04502/3 (left panel) complemented *in trans* with BACOVA\_04502, BACOVA\_04503, BACOVA\_04502/3 or vector alone (pFD340) and *Bo*  $\Delta$ 04502 (right panel) with BACOVA\_04502 or vector alone grown in defined inulin media. See Supplementary Information Fig. 1 for uncropped scanned images. S, stationary phase.



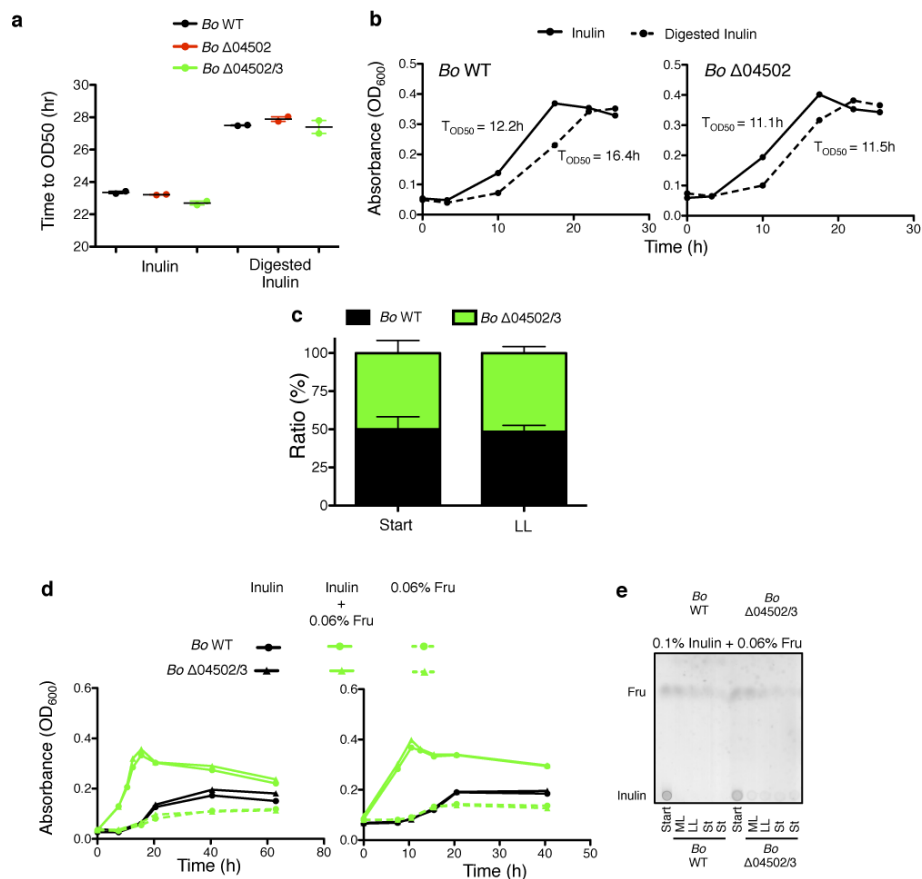


**Extended Data Figure 3 | *SusC* and *SusD* homologues BACOVA\_04505 and BACOVA\_04505 are required for inulin utilization.** **a**, Independent experiments for Fig. 2e. Left panel  $n = 2$  biological replicates, right panel each line represents  $n = 1$  sample per condition. **b**, Complementation of  $Bo$   $\Delta$ 04504,  $Bo$   $\Delta$ 04505 and  $\Delta$ 04504/5 mutants with the genes *in trans*. Growth of  $Bo$   $\Delta$ 04504,  $Bo$   $\Delta$ 04505,  $Bo$   $\Delta$ 04504/5 with BACOVA\_04504, BACOVA\_04505, BACOVA\_04504/5 or vector alone (pFD340) *in trans* in defined inulin media. Each line represents  $n = 1$  sample per condition. In all panels, error bars represent standard error;  $P$  values derived from two-tailed Student's  $t$ -test.



**Extended Data Figure 4 | Costs of extracellular inulin digestion by *B. ovatus*.** **a, b**, Independent experiments for Fig. 3a, c.  $n = 3$  biological replicates at day 1, 2;  $n = 2$ , biological replicates at day 4 (**a**). In upper and upper middle panels biological replicates of each condition are plotted as individual lines ( $n = 2$  cell culture biological replicates); in lower middle and lower panels each line represents  $n = 1$  sample per condition (**b**).

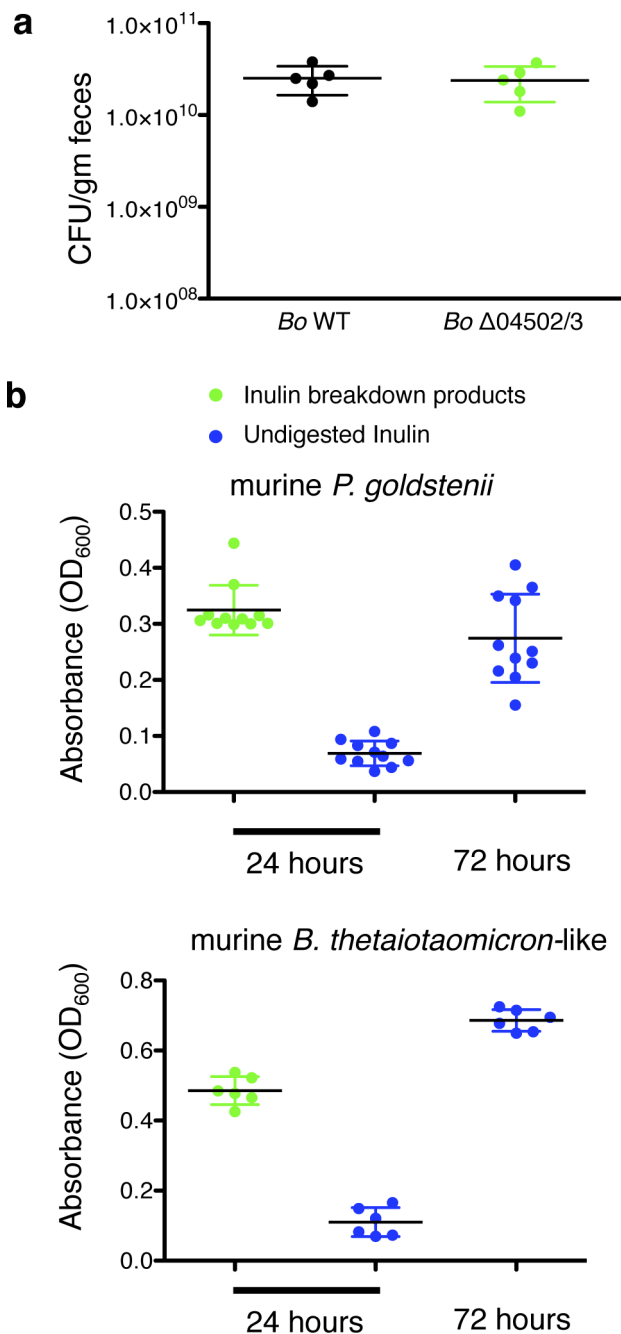
**c**, TLC analysis of conditioned media from *Bo* WT and *Bo*  $\Delta$ 04502/3 cultured in 0.5% inulin with trace (0.06%) amounts of fructose. St, stationary phase. See Supplementary Information Fig. 1 for uncropped scanned images. In all panels, error bars represent standard error;  $P$  values derived from two-tailed Student's  $t$ -test.



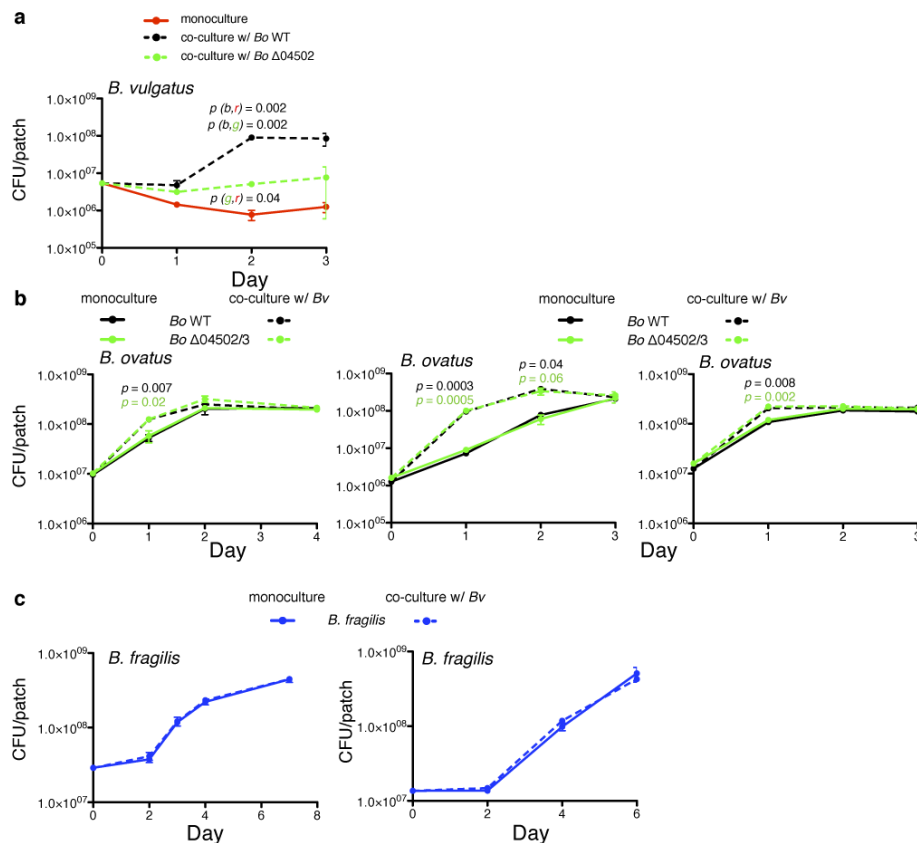
**Extended Data Figure 5 | Preferential utilization of undigested inulin by *B. ovatus* and costs of inulin digestion by 04502/3.** **a**, Time to mid-log (estimated at 50% maximal OD (OD<sub>50</sub>) by linear regression analysis) of Fig. 3d. **b**, Additional independent experiments for Fig. 3d. Each line represents  $n = 1$  sample per condition. **c**, Competition of *Bo* WT and *Bo*  $\Delta$ 04502/3 co-cultured in 0.5% inulin with trace (0.06%) amounts of fructose.  $n = 2$  cell culture biological replicates; **d**, **e**, Growth (**d**) and TLC

analysis of conditioned media (**e**) of *Bo* WT and *Bo*  $\Delta$ 04502/3 cultured in 0.1% inulin with trace (0.06%) amounts of fructose. See Supplementary Information Fig. 1 for uncropped scanned images. For **d**, each panel is an independent experiment. Each condition is plotted as individual lines in each panel ( $n = 2$  cell culture biological replicates). In all panels, error bars represent standard error;  $P$  values derived from two-tailed Student's  $t$ -test.



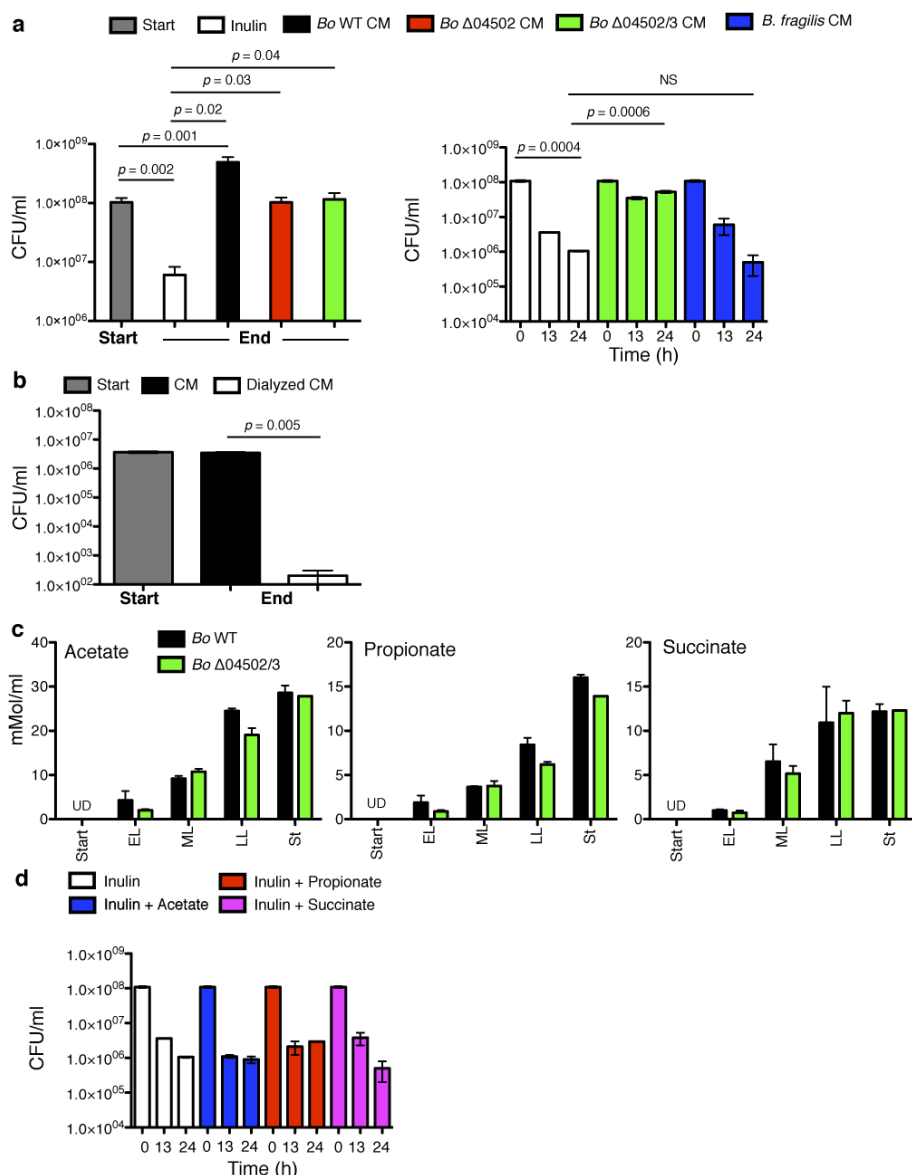


**Extended Data Figure 6 | A complex mouse microbiota differentially affects *B. ovatus* WT and  $\Delta 04502/3$  pre-colonized gnotobiotic mice and analysis of cross feeding of the predominant murine Bacteroidales of the murine gut microbiota. a,** Germ-free mice were monocolonized with *Bo* WT or  $\Delta 04502/3$  and maintained on a diet supplemented with inulin as the sole polysaccharide and housed under gnotobiotic conditions for 2 weeks. Bacteria were enumerated from faeces before gavage with intestinal microbiota of conventionally raised mice ( $n = 5$  mice, cell culture biological replicates). **b,** Growth of two dominant mouse microbiota Bacteroidales strains (*Parabacteroides goldsteinii* and a strain with 96% 16S rRNA gene identity to *B. thetaiotaomicron*) with inulin breakdown products derived from the conditioned media of *B. ovatus* grown in inulin (all inulin had been digested) or undigested inulin minimal media. Each data point is a different isolate of the indicated species from the caeca of the conventionally raised mice used for gavage ( $n = 11$  isolates for upper panel,  $n = 6$  for lower panel). In all panels, error bars represent standard error.



**Extended Data Figure 7 | The *B. ovatus*, but not *B. fragilis* benefits from *B. vulgatus* in co-culture in inulin.** **a**, **b**, Independent experiments for Fig. 4b, d. The left panel corresponds to starting culture with  $\sim 10^7$  CFU *Bo* corresponding to starting culture of  $\sim 10^6$  CFU *Bo* in Fig. 4d. The two right panels are a duplicate pair of experiments of starting CFU of  $10^6$  and  $10^7$ . In each panel  $n = 2$  biological replicates. **c**, Enumeration of *B. fragilis* in monoculture or co-culture with *Bv* on defined inulin plates,

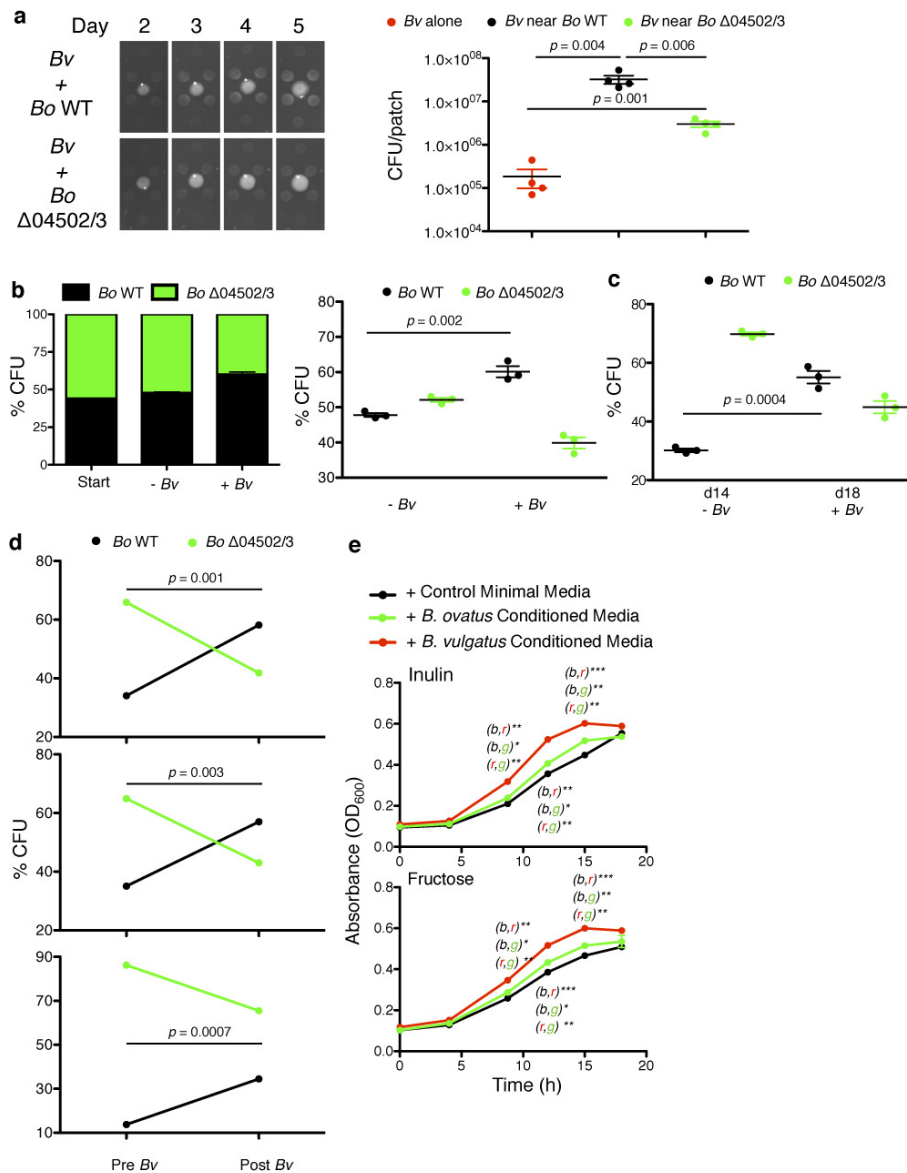
$n = 3$  cell culture biological replicates. Letters in parentheses refer to values correlating to colour of line used for statistical analysis. In **a**, for example,  $p(g,r)$  refers to comparison of values of green (*g*) and red (*r*) values at the time-point indicated. In **b**, colour of *P* value indicates comparison of monoculture and co-culture for the given condition at the time-point indicated. For all panels, error bars represent standard error; *P* values derived from two-tailed Student's *t*-test.



**Extended Data Figure 8 | Secreted factors from *Bo* and isogenic mutants, but not *B. fragilis* (*Bf*), support *Bv* survival.** **a**, Growth of *Bv* in conditioned media from *Bo* WT,  $\Delta$ 04502,  $\Delta$ 04502/3 or *Bf* grown in defined media with inulin as the sole carbohydrate and in inulin media. End time-point corresponded to peak growth of *Bv* in conditioned media derived from *Bo* WT. *B. fragilis*, which utilizes inulin<sup>1,3</sup> but similar to *Bo*  $\Delta$ 04502/3 does not liberate inulin breakdown products<sup>3</sup>, does not support the survival of *Bv* during co-culture. Left and right panels are independent experiments. Left panel; start,  $n = 4$  biological replicates; end,  $n = 2$  biological replicates. Right panel;  $t_0$ ,  $n = 4$  cell culture biological replicates;  $t_{13}$  and 24,  $n = 2$  biological replicates. **b**, Growth of *Bv* in dialysed (2 kDa MW membrane) or undialysed conditioned media from

*Bo*  $\Delta$ 04502/3 grown in inulin,  $n = 2$  cell culture biological replicates. **c**, Gas chromatographic analysis of acetate, propionate and succinate in conditioned media during growth of *Bo* WT and  $\Delta$ 04502/3 in defined media with inulin as the sole carbohydrate. Other volatile and non-volatile substances (as listed in Methods) were undetectable,  $n = 2$  cell culture biological replicates, except  $\Delta$ 04502/3 stationary phase,  $n = 1$ . **d**, Growth of *Bv* in defined medium with inulin as the sole carbohydrate with or without addition with 15 mM of acetate, propionate or succinate,  $t = 0$ ,  $n = 4$  biological replicates;  $t = 13$  and 24 h,  $n = 2$  biological replicates. CM, conditioned media; UD, undetectable. For all panels, error bars represent standard error;  $P$  values derived from two-tailed Student's  $t$ -test.

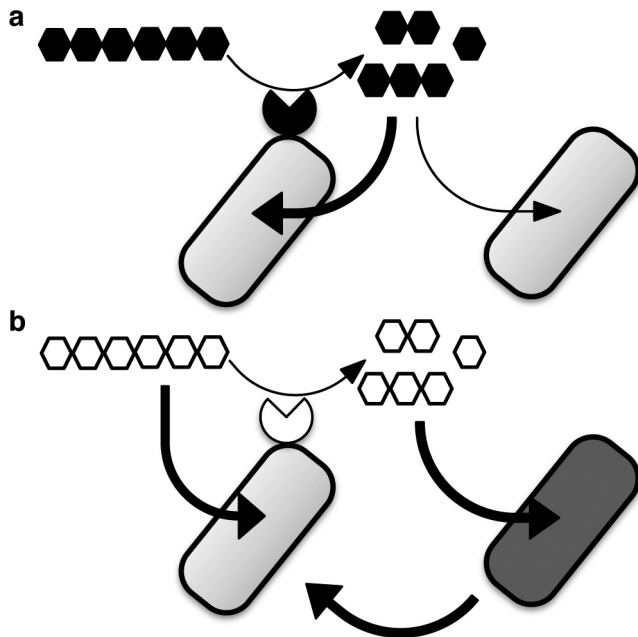




### Extended Data Figure 9 | Spatial aspects of mutualism between *Bo* and *Bv* via cross feeding and interspecies cooperation *in vivo* via 04502/3.

**a**, Independent experiments for Fig. 4c,  $n = 4$  biological replicates. **b**, Independent experiments for Fig. 4e,  $n = 3$  cell culture biological replicates. **c**, Scatter plot of experiment in Fig. 4f,  $n = 3$  cell culture biological replicates. **d**, Ratios of wild type and Δ04502/3 in faeces 21 days after co-colonization of three germ-free mice ( $n = 3$  mice biological replicates) on a diet of inulin as the sole dietary polysaccharide (pre-*Bv*) and then four days after introduction of *Bv* (post-*Bv*). Each panel shows the ratio pre- and post- *Bv* of an individual mouse.  $P$  values are Fisher exact test comparing the frequency of *Bo* WT and Δ04502/3 pre- and post- colonization with *Bv* for each individual mouse. At day 21, all mice were colonized with a higher ratio of the mutant (ranging to 86%

in the mouse depicted in the lowest panel), with each mouse showing a statistically significant increase in the proportion of the wild type after introduction of *Bv*. **e**, Growth of *B. ovatus* in 0.5% inulin (upper panel) or fructose (lower panel) in minimal media to which filter sterilized conditioned media from early log,  $A_{600\text{ nm}}$  matched growth of *Bv* or *Bo* in 0.125% fructose minimal media or fresh 0.125% fructose minimal media control was added at 1:1 ratio. In **e**, numbers refer to  $P$  values (\* $P < 0.05$ , \*\* $P < 0.01$ , \*\*\* $P < 0.001$ ) of comparison of values of green (g), red (r) or black (b) values by unpaired, two-tailed student  $t$ -test at the time-point indicated,  $n = 2$  cell culture biological replicates. For all panels, error bars represent standard error; for all panels except **d**,  $P$  values derived from two-tailed Student's  $t$ -test.



**Extended Data Figure 10 | Schematic of forms of cooperativity via polysaccharide digestion among Bacteroidales. a.** Limited cooperation. Privatization of extracellularly digested public goods by the individual performing the digestion leads to greater individual (thick arrow) than shared benefits (thin arrow) as seen in *Bt* during growth on levan and amylopectin. **b.** Cooperation between species is seen between *Bo* and *Bv* during growth on inulin. Surface digestion of inulin by *Bo* creates breakdown products that it does not need to grow on inulin. Rather, inulin breakdown represents a dedicated cross-feeding system that provides benefits to *Bv*, with reciprocal fitness benefits to *Bo*.

# Molecular mechanism of APC/C activation by mitotic phosphorylation

Suyang Zhang<sup>1\*</sup>, Leifu Chang<sup>1\*</sup>, Claudio Alfieri<sup>1</sup>, Ziguo Zhang<sup>1</sup>, Jing Yang<sup>1</sup>, Sarah Maslen<sup>1</sup>, Mark Skehel<sup>1</sup> & David Barford<sup>1</sup>

In eukaryotes, the anaphase-promoting complex (APC/C, also known as the cyclosome) regulates the ubiquitin-dependent proteolysis of specific cell-cycle proteins to coordinate chromosome segregation in mitosis and entry into the G1 phase<sup>1,2</sup>. The catalytic activity of the APC/C and its ability to specify the destruction of particular proteins at different phases of the cell cycle are controlled by its interaction with two structurally related coactivator subunits, Cdc20 and Cdh1. Coactivators recognize substrate degrons<sup>3</sup>, and enhance the affinity of the APC/C for its cognate E2 (refs 4–6). During mitosis, cyclin-dependent kinase (Cdk) and polo-like kinase (Plk) control Cdc20- and Cdh1-mediated activation of the APC/C. Hyperphosphorylation of APC/C subunits, notably Apc1 and Apc3, is required for Cdc20 to activate the APC/C<sup>7–12</sup>, whereas phosphorylation of Cdh1 prevents its association with the APC/C<sup>9,13,14</sup>. Since both coactivators associate with the APC/C through their common C-box<sup>15</sup> and Ile-Arg tail motifs<sup>16,17</sup>, the mechanism underlying this differential regulation is unclear, as is the role of specific APC/C phosphorylation sites. Here, using cryo-electron microscopy and biochemical analysis, we define the molecular basis of how phosphorylation of human APC/C allows for its control by Cdc20. An auto-inhibitory segment of Apc1 acts as a molecular switch that in apo unphosphorylated APC/C interacts with the C-box binding site and obstructs engagement of Cdc20. Phosphorylation of the auto-inhibitory segment displaces it from the C-box-binding site. Efficient phosphorylation of the auto-inhibitory segment, and thus relief of auto-inhibition, requires the recruitment of Cdk–cyclin in complex with a Cdk regulatory subunit (Cks) to a hyperphosphorylated loop of Apc3. We also find that the small-molecule inhibitor, tosyl-L-arginine methyl ester, preferentially suppresses APC/C<sup>Cdc20</sup> rather than APC/C<sup>Cdh1</sup>, and interacts with the binding sites of both the C-box and Ile-Arg tail motifs. Our results reveal the mechanism for the regulation of mitotic APC/C by phosphorylation and provide a rationale for the development of selective inhibitors of this state.

To understand how multi-site phosphorylation of numerous APC/C subunits stimulates the capacity of Cdc20 to control the APC/C we determined a series of APC/C structures in different functional states to near-atomic resolution (Extended Data Table 1a). We used the kinases Cdk2–cyclin A3–Cks2 and Polo (Plk1) to phosphorylate *in vitro* recombinant human APC/C<sup>11,12,18,19</sup> (Extended Data Fig. 1a), obtaining APC/C in the mitotic state that can be activated by Cdc20 (Extended Data Fig. 1b, lanes 9, 10). This reconstituted APC/C recapitulates Cdk- and Plk1-dependent activation of endogenous APC/C<sup>Cdc20</sup> (refs 8–12). Kinase treatment resulted in a complete upshift of the Apc3 subunit as visualized on SDS–PAGE, indicative of stoichiometric phosphorylation (Extended Data Fig. 1a, c). Almost 150 phosphorylation sites were identified in phospho-APC/C by mass spectrometry (Extended Data Tables 2 and 3), matching published data<sup>12,20–22</sup>. These sites lie within disordered regions of the APC/C<sup>23</sup>. Incubating the APC/C with both Cdk2–cyclin A3–Cks2 and Plk1 simultaneously was necessary to

obtain full activation (Extended Data Fig. 1b). Consistent with ref. 12, treatment with Cdk2–cyclin A3–Cks2 alone resulted in lower APC/C activation, whereas phosphorylation with Plk1 alone did not activate the APC/C.

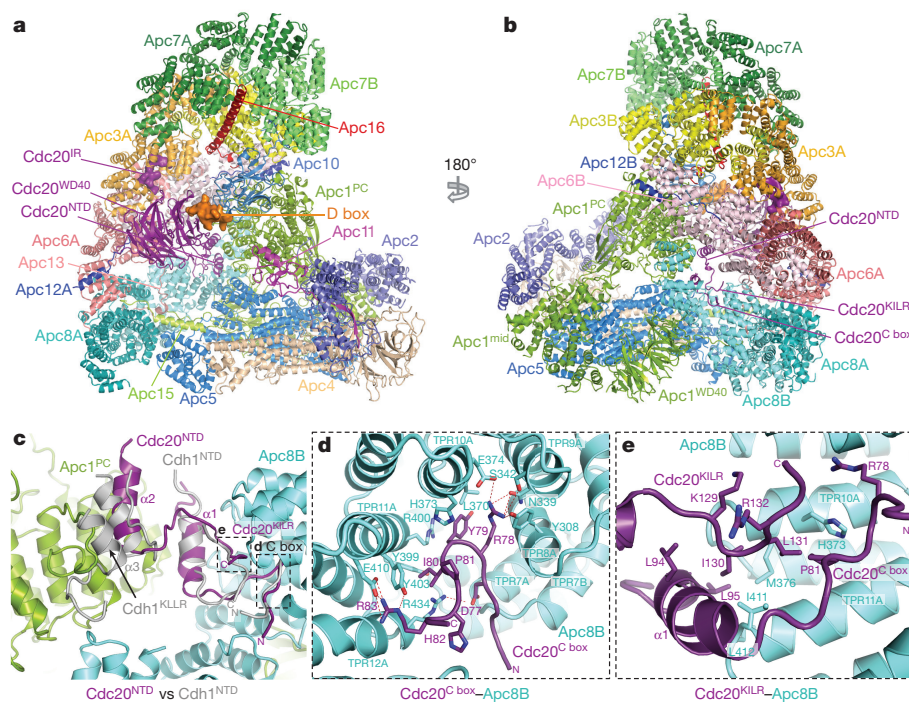
To gain insights into the molecular interactions between Cdc20 and mitotic APC/C, a ternary complex was assembled using phosphorylated APC/C, Cdc20 and a high-affinity substrate Hsl1 (APC/C<sup>Cdc20-Hsl1</sup>) (Extended Data Fig. 1d) for cryo-EM analysis (Fig. 1a, b, Extended Data Fig. 1e–g and Extended Data Table 1). Owing to the low occupancy of Cdc20 bound to the APC/C, combined with conformational heterogeneity, only 9% of the particles were used for the final reconstruction of APC/C<sup>Cdc20-Hsl1</sup>, with the remainder being in either the apo state (72%) or in a hybrid state (Extended Data Fig. 2). Most static regions of the complex extend beyond 3.9 Å resolution, whereas the catalytic module (Apc2 and Apc11), as well as the WD40 domain of Cdc20, are more flexible (Extended Data Fig. 3a).

APC/C<sup>Cdc20-Hsl1</sup> adopts an active conformation with the catalytic module in the upward position reminiscent of APC/C<sup>Cdh1-Hsl1</sup> (ref. 6) (Fig. 1a, b, Extended Data Fig. 4a, b and Supplementary Video 1) and in agreement with a low-resolution negative-stain EM reconstruction of APC/C<sup>Cdc20</sup> (ref. 24). Relative to Cdh1<sup>WD40</sup>, the Cdc20<sup>WD40</sup> domain is shifted away from the APC/C by as much as 10 Å (Extended Data Fig. 3b). Its interaction with the APC/C involves only its N-terminal domain (Cdc20<sup>NTD</sup>) and the C-terminal Ile–Arg (IR) tail (Cdc20<sup>IR</sup>) (Fig. 1c–e, Extended Data Figs 3c–e and 4c, d). Compared with Cdh1<sup>NTD</sup>, Cdc20<sup>NTD</sup> forms fewer contacts with both Apc1 and Apc8B (Fig. 1c–e). However, the crucial C-box motif (DRYIPxR) represents a structurally conserved region common to both coactivators (Fig. 1d, Extended Data Figs 3c, d and 4d). Cdc20<sup>Cbox</sup> forms a network of electrostatic interactions with Apc8B, centred on the crucial Arg78 (ref. 23), and augmented by non-polar interactions involving its Tyr79 and Ile80 residues (Fig. 1d). A KILR motif also present within Cdc20<sup>NTD</sup> is essential for Cdc20 association with the APC/C (ref. 25) and the APC/C<sup>Cdc20-Hsl1</sup> structure reveals that Ile130 and Leu131 of Cdc20<sup>KILR</sup> are inserted into a hydrophobic pocket of the TPR (tetratricopeptide repeat) superhelix of Apc8B, further stabilizing the conformation of the C box (Fig. 1e). Similar C-box stabilization is present in Cdh1<sup>NTD</sup>, but is instead provided by a loop structurally unrelated to Cdc20<sup>KILR</sup> (Fig. 1c)<sup>23</sup>. By contrast, the KLLR-motif of Cdh1<sup>NTD</sup> is located in a leucine-zipper-like  $\alpha$ -helix ( $\alpha$ 3) that forms a hydrophobic interface with Apc1 (ref. 23) (Fig. 1c and Extended Data Fig. 4d). The absence of an equivalent to the Cdh1  $\alpha$ 3 helix in Cdc20 suggests a weaker mode of binding of Cdc20<sup>NTD</sup> relative to Cdh1<sup>NTD</sup>.

EM density for Cdc20<sup>IR</sup> is weaker than for Cdh1<sup>IR</sup> and it lacks the associated  $\alpha$ -helix of Cdh1<sup>IR</sup> (Extended Data Fig. 3e). This could account for the lower affinity of the APC/C for Cdc20<sup>IR</sup> compared with Cdh1<sup>IR</sup> (ref. 17). Nonetheless, the crucial Ile–Arg interaction of Cdc20<sup>IR</sup> with the TPR superhelix of Apc3A is conserved between the two coactivators (Extended Data Fig. 3e). Importantly, because in the APC/C<sup>Cdc20-Hsl1</sup> EM structure, densities corresponding to phosphorylated residues

<sup>1</sup>MRC Laboratory of Molecular Biology, Francis Crick Avenue, Cambridge CB2 0QH, UK.

\*These authors contributed equally to this work.



**Figure 1 | EM reconstructions of the APC/C<sup>Cdc20-Hsl1</sup> complex and comparison of Cdc20<sup>NTD</sup> and Cdh1<sup>NTD</sup>.** **a, b**, Two views of APC/C<sup>Cdc20-Hsl1</sup> shown in cartoon with the D-box and Cdc20<sup>IR</sup> highlighted in surface representation. Cdc20 binds to the APC/C in juxtaposition to Apc10 to form the substrate recognition module. Apc1 is composed of an N-terminal WD40 domain Apc1<sup>WD40</sup>, a middle domain Apc1<sup>mid</sup> and a C-terminal Apc1<sup>PC</sup> domain. Apc11 is modelled based on the APC/C<sup>Cdh1-Emi1</sup> structure (PDB 4UI9)<sup>22</sup>. **c**, Both Cdc20<sup>NTD</sup> (purple) and Cdh1<sup>NTD</sup> (grey, aligned to APC/C<sup>Cdc20-Hsl1</sup>)<sup>23</sup> interact with Apc1 and Apc8B, whereas Cdh1<sup>NTD</sup> contains an additional  $\alpha$ 3 helix associating with Apc1. **d**, The crucial C-box motif is well conserved between the two coactivators and forms extensive interactions with Apc8B. **e**, The KLLR motif of Cdh1 is present in the  $\alpha$ 3 helix to engage Apc1, whereas the related Cdc20<sup>KLLR</sup> motif contacts Apc8B to augment C-box binding.

are not visible, we find no evidence that phosphorylated regions of the APC/C either directly or indirectly contact Cdc20. This suggested that APC/C phosphorylation invokes a conformational change of apo APC/C that promotes its association with Cdc20.

To explore this possibility, we determined cryo-EM structures of apo APC/C in both the unphosphorylated and phosphorylated states at near-atomic resolution (Fig. 2a, b, Extended Data Fig. 5a, b, Extended Data Table 1 and Supplementary Video 1). In both states, the catalytic module adopts an inactive conformation (Extended Data Fig. 4a, b) as seen in the previous 8 Å resolution reconstruction<sup>6</sup>. However, three-dimensional classification of the atomic resolution EM maps of both apo states showed that the majority of Apc3A adopts a closed conformation resembling the Apc3 crystal structure in which an  $\alpha$ -helix (TPR12A) occupies and blocks the IR tail-binding pocket (Extended Data Fig. 5c)<sup>26</sup>. In ~30% of particles Apc3A adopts an open conformation identical to the IR tail-bound state. Thus inter-conversion of Apc3A between closed and open IR tail-accessible states is not controlled by phosphorylation.

Phosphorylated and unphosphorylated apo APC/C EM maps are very similar in structure (Extended Data Fig. 5a, b), except for a notable difference in the region of the C-box binding site (Fig. 2c–f). In unphosphorylated APC/C, an unassigned segment of EM density of ~15 residues indicative of an elongated loop connected to a short  $\alpha$ -helix is located at the C-box binding pocket of Apc8B (Fig. 2c–e). The equivalent EM density is not present in phosphorylated APC/C (Fig. 2f). The WD40 domain of Apc1, positioned in close proximity to this density, incorporates two highly phosphorylated regions (residues 307–395 (300s loop) and residues 515–579 (500s loop)) (Fig. 2c, d and Extended Data Table 2), which have not previously been assigned in either APC/C<sup>Cdc20-Hsl1</sup> or APC/C<sup>Cdh1-Emi1</sup> structures<sup>23</sup>. The 300s loop would be predicted to project towards the C-box binding site of Apc8B (Fig. 2c, d and Extended Data Fig. 5d), implicating it as a candidate for the unassigned density segment.

We determined the structure of an APC/C mutant with the 300s loop of Apc1 deleted (APC/C <sup>$\Delta$ Apc1-300s</sup>) (Fig. 2g and Extended Data Table 1a). In this structure, the C-box binding pocket of APC/C <sup>$\Delta$ Apc1-300s</sup> is devoid of EM density even without *in vitro* phosphorylation (Fig. 2g), consistent with its assignment to the 300s loop. Furthermore, ubiquitination assays showed that APC/C <sup>$\Delta$ Apc1-300s</sup> was constitutively activated by Cdc20 and that phosphorylation did not enhance its activity (Fig. 3a,

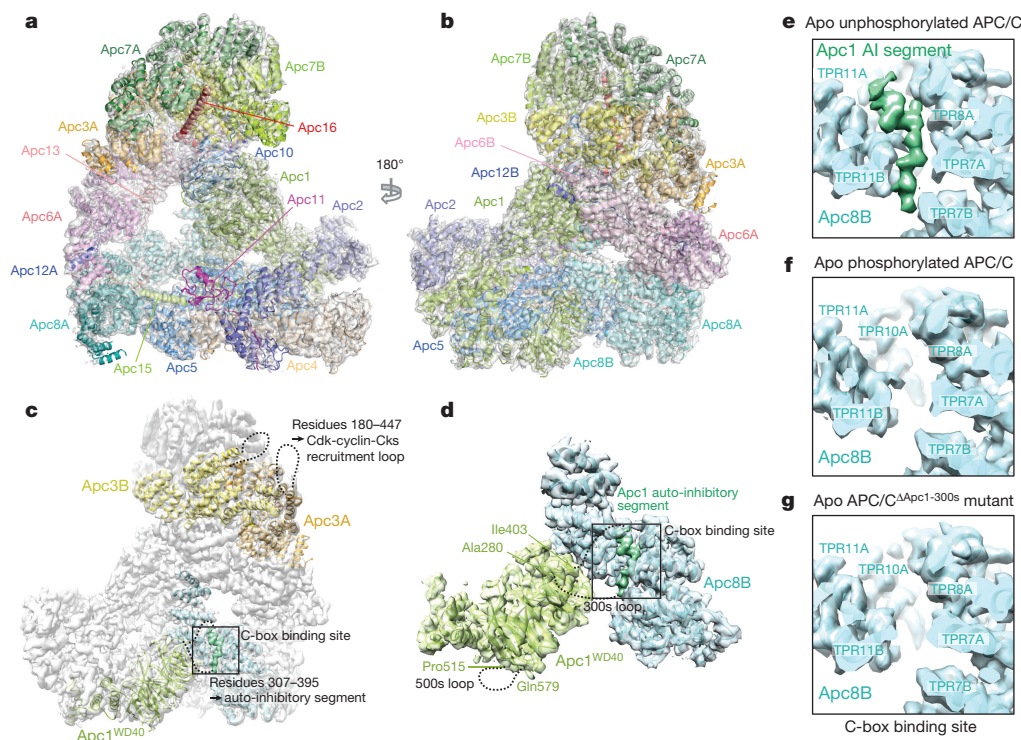
compare lanes 6, 7 to 8, 9). This indicates that unphosphorylated APC/C is maintained in an auto-inhibited conformation by an Apc1 auto-inhibitory segment, present within the 300s loop that sterically impedes the C-box site from binding Cdc20. In support of this idea, analytical size-exclusion chromatography showed that Cdc20 forms a binary complex with phosphorylated APC/C, but not with unphosphorylated APC/C (Extended Data Fig. 6a), in agreement with refs 10, 12.

To identify the auto-inhibitory segment within the Apc1 300s loop, we synthesized a set of eight overlapping peptides of 20 residues, spanning the 300s loop, and tested their potential to inhibit APC/C <sup>$\Delta$ Apc1-300s</sup> (Fig. 3b). Notably, peptide 7 (residues 361–380, Fig. 3b, lane 9) potently suppressed Cdc20-dependent APC/C <sup>$\Delta$ Apc1-300s</sup> activity. Interestingly, an Arg368–Phe369 pair of this peptide, that resembles the Arg78–Tyr79 of the C box, is flanked by four serine residues phosphorylated in mitotic APC/C (Fig. 3c and Extended Data Table 2)<sup>12,21,22</sup>.

EM density for the auto-inhibitory segment is weak, probably due to partial occupancy at the C-box site (Fig. 3c). Nevertheless, side chain density similar to Arg78–Tyr79 of the C box suggests a fit for the Arg–Phe of peptide 7 (Arg368–Phe369) (Fig. 3c). To test the possibility that the Apc1 auto-inhibitory segment corresponds to peptide 7, we synthesized mutants of peptide 7 and also introduced the equivalent mutations into Apc1 of the recombinant APC/C. Replacing Arg368 with Glu in Apc1 resulted in a Cdc20-dependent activation of unphosphorylated APC/C (Fig. 3d), and reduced the inhibitory potency of peptide 7 towards APC/C <sup>$\Delta$ Apc1-300s</sup> (Extended Data Fig. 6c). A similar result was obtained on substituting glutamates for the four neighbouring serine residues (Ser364, Ser372, Ser373 and Ser377 (APC/C <sup>$\Delta$ Apc1-4S/E</sup>)) to mimic phosphorylation (Fig. 3d and Extended Data Fig. 6c). Phosphorylation of Ser377 of peptide 7 relieved the inhibition only partially. These findings suggest that Arg368 anchors the Apc1 auto-inhibitory segment to the C-box binding site, mimicking the Cdc20 C box, and maintaining the apo APC/C in an auto-inhibited state. Phosphorylation of the four neighbouring serine residues would destabilize its association with the C-box site (Figs 3c and 4a). This mechanism could be further tested in an *in vivo* context by cellular assays.

Our results so far reveal that the critical determinant of APC/C<sup>Cdc20</sup> activation by phosphorylation is displacement of the Apc1 auto-inhibitory segment from the C-box site. However, since Apc3 is hyperphosphorylated in mitosis, and Cks stimulates both Cdk-dependent activation of APC/C<sup>Cdc20</sup> (refs 18, 19) and Apc1 and Apc3





**Figure 2 | Apo unphosphorylated APC/C is repressed by an Apc1 auto-inhibitory segment.** **a, b**, Two views of the phosphorylated apo APC/C structure in cartoon within the 3.4 Å EM map (grey). The catalytic module (Apc2 and Apc11) is in the inactive conformation. **c, d**, EM map of unphosphorylated apo APC/C. Apc1 has two highly phosphorylated loops within its WD40 domain (green). Whereas the 300s loop (residues 307–395) is pointing towards the Apc1 auto-inhibitory segment density (dark green) at the C-box binding site (black box), the 500s loop is facing in the opposite direction. The hyperphosphorylated Apc3 loop

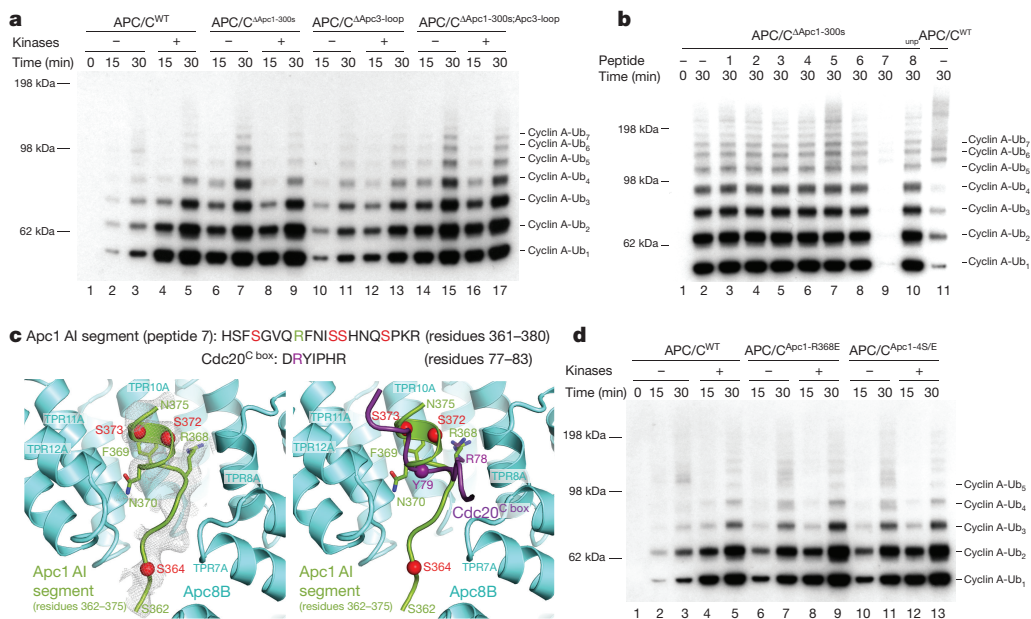
phosphorylation<sup>18</sup>, and interacts with Apc3 (refs 18,27,28), we tested whether Apc3 phosphorylation also had a role in APC/C activation. In Apc3 about 50 phosphorylation sites are clustered in a large disordered loop comprising residues 180–450 (Extended Data Table 2) located on the same face of the APC/C as the Apc1 300s loop (Fig. 2c). In contrast to APC/C $\Delta$ Apc1-300s, instead of stimulating unphosphorylated APC/C, deletion of the Apc3 loop (APC/C $\Delta$ Apc3-loop) reduced the phosphorylation-mediated activation of APC/C (Fig. 3a, lanes 10–13). Similarly, Cdk2–cyclin A3 failed to stimulate the APC/C activity in the absence of Cks2 (Extended Data Fig. 1b). However, combining deletions of both the Apc3 and Apc1 300s loops (APC/C $\Delta$ Apc1-300s;Apc3-loop) restored activity to that of wild-type phosphorylated APC/C (Fig. 3a, lanes 14, 15). Since deletion of the Apc3 loop disrupts APC/C association with Cdk–cyclin–Cks (Extended Data Fig. 1c, lanes 6, 8), a likely explanation for our results and for the lag phase that accompanies APC/C activation by Cdk1–cyclin B–Cks<sup>19</sup>, is that Apc3 phosphorylation recruits Cdk–cyclin–Cks through Cks<sup>18,27,28</sup> to stimulate Apc1 auto-inhibitory segment phosphorylation. Cdk–cyclin–Cks association with the Apc3 loop would allow for a kinetically more efficient intra-molecular phosphorylation of the Apc1 auto-inhibitory segment that only becomes accessible to Cdk when transiently displaced from the C-box site (Figs 3c and 4a).

To determine whether Apc3 loop-mediated interactions with the Cks2 subunit facilitated Apc1 300s loop phosphorylation, we analysed phosphorylation of the Apc1 300s loop in conditions where such interactions are disrupted. Either deletion of the Apc3 loop or omission of the Cks2 subunit from the phosphorylation reaction, conditions that reduce APC/C activation (Fig. 3a and Extended Data Fig. 1b), resulted in the same reduction of Apc1 300s loop phosphorylation (Extended Data Table 2). Specifically, mitotic phospho-sites associated with relief

(residues 180–447) is located at the back of the APC/C and functions as a Cdk recruitment site. The views in **b** and **c** are similar to Fig. 1b. **e–g**, Close-up views of the C-box binding site in the EM maps of apo unphosphorylated and phosphorylated APC/C and an APC/C $\Delta$ Apc1-300s mutant with the Apc1 300s loop deleted. An elongated loop density (dark green) for the Apc1 auto-inhibitory (AI) segment was observed in the apo unphosphorylated state (**e**), but the density is absent in apo phosphorylated APC/C (**f**). Deletion of the Apc1 300s loop shows a similar loss of C-box site-associated density (**g**).

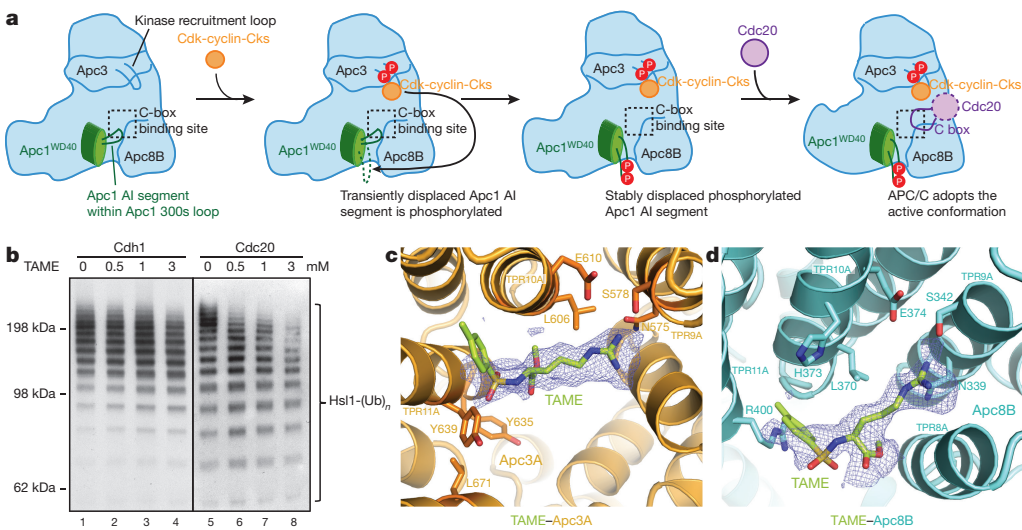
of auto-inhibition (Ser364, Ser372 and Ser373) (Extended Data Table 2, columns 3 and 8) are not modified when Cdk targeting to Apc3 is disrupted (Extended Data Table 2, columns 4 and 5). In agreement with<sup>12,22</sup> Ser362 and Ser364 are phosphorylated by Plk1 and Cdk2–cyclin A3–Cks2, respectively, indicating that Ser364 phosphorylation confers partial activation, whereas phosphorylation of Ser372 and Ser373 requires the presence of both kinases, possibly owing to inter-dependent priming reactions. Since the stimulatory phospho-Ser364 of the auto-inhibitory segment (Fig. 3c) is a non Cdk-consensus site, the relaxed specificity of Cdk2–cyclin A3–Cks2 phosphorylation of this site may be conferred through targeting of the kinase to the APC/C through the Apc3 loop.

Cdc20 and Cdh1 bind to common sites on the APC/C yet only Cdc20 association is regulated by APC/C phosphorylation. In agreement with ref. 9, we find that Cdh1 activates unphosphorylated and phosphorylated APC/C to a similar extent (Extended Data Fig. 6d, e). Comparing unphosphorylated APC/C and APC/C $\Delta$ Apc1-300s we detect some stimulation with APC/C $\Delta$ Apc1-300s at low Cdh1 concentrations (Extended Data Fig. 6e). This is consistent with the enhancement of Cdh1 binding to phosphorylated APC/C and APC/C $\Delta$ Apc1-300s (Extended Data Fig. 6b). Testing the phosphorylation-dependent activity of a set of Cdc20–Cdh1 chimaeras indicated that the more extensive interactions involving Cdh1 and the APC/C, mediated by both Cdh1<sup>NTD</sup> and Cdh1<sup>IR</sup> (Extended Data Fig. 6d), with a contribution from the Cdh1<sup>NTD</sup>  $\alpha$ 3 helix (Fig. 1c and Extended Data Fig. 6f), account for the capacity of Cdh1 to activate unphosphorylated APC/C. Although Cdh1 association would also require displacement of the Apc1 auto-inhibitory segment, a potential mechanism to explain the dependency of Cdc20 on APC/C phosphorylation is that the higher affinity of Cdh1 for the



**Figure 3 | The Apc1 auto-inhibitory segment binds to the C-box binding site and mimics the Cdc20<sup>Cbox</sup>.** **a**, A native substrate Cdk2–cyclin A2–Cks2 was used for ubiquitination assays. *In vitro* phosphorylated APC/C (both Cdk2–cyclin A3–Cks2 and Plk1) can be activated by Cdc20 (lanes 1–5). Deletion of the Apc1 300s loop activated the APC/C without phosphorylation (lanes 6, 7), and kinase treatment of APC/C<sup>ΔApc1-300s</sup> did not enhance APC/C activity. The APC/C<sup>ΔApc3-loop</sup> mutant showed similar activity as unphosphorylated APC/C (lanes 10, 11 versus 2, 3 and 4, 5), but had reduced activation by phosphorylation. Nevertheless, deletion of both Apc1 300s and Apc3 loops (APC/C<sup>ΔApc1-300s;Apc3-loop</sup>) restored activity to that of wild-type (WT) phosphorylated APC/C and unphosphorylated APC/C<sup>ΔApc1-300s</sup> (lanes 14–17). **b**, Identification of the Apc1 auto-inhibitory segment occupying the C-box binding site by assessing the inhibitory effect of eight peptides spanning the Apc1 300s loop. A single peptide (peptide 7, residues 361–380) suppressed the activity of

APC/C<sup>ΔApc1-300s</sup> (lane 9), indicating that this peptide blocks the C-box binding site. A control with wild-type unphosphorylated APC/C (unp-APC/C<sup>WT</sup>) is in lane 11. **c**, The Apc1 auto-inhibitory (AI) segment (peptide 7, residues 361–380) shares sequence similarity with Cdc20<sup>C box</sup>. A model for the auto-inhibitory segment (green) was fitted into the EM density of the apo unphosphorylated APC/C map (grey). Arg368 overlaps with the crucial Arg78 of Cdc20<sup>C box</sup> (purple, right panel). The flanking serines shown to be phosphorylated are highlighted as red spheres. Ser377 is outside the observed EM density. **d**, Mutation of a single Arg368 residue (APC/C<sup>Apc1-R368E</sup>) or mutating its four neighbouring serine residues (Ser364, Ser372, Ser373, Ser377) to glutamates (APC/C<sup>Apc1-4S/E</sup>) activated the APC/C without phosphorylation. 30 nM Cdc20 was used for the assay in **a** and 20 nM Cdc20 for the assays in **b** and **d**. Experiments in **a** and **d** were replicated three times and in **b** five times. See Supplementary Fig. 1 for gel source data.



**Figure 4 | Mechanism for APC/C activation by mitotic phosphorylation and the molecular basis for TAME inhibition.** **a**, Cartoon showing the mechanism of APC/C activation by Apc1 and Apc3 phosphorylation-induced relief of auto-inhibition. Artificial relief of auto-inhibition, either by deletion of the Apc1 auto-inhibitory (AI) segment or by its phospho-mimicking mutants, obviates the need to phosphorylate Apc3. **b**, TAME has only a small inhibitory effect on APC/C<sup>Cdh1</sup> (lanes 1–4), whereas it markedly reduced APC/C<sup>Cdc20</sup> activity (lanes 5–8). The activity assay was performed with phosphorylated

wild-type (WT) APC/C and substrate Hsl1 at a coactivator concentration of 10 nM. This experiment was replicated three times. See Supplementary Fig. 1 for gel source data. **c, d**, EM reconstruction of APC/C<sup>ΔAPC1-300s</sup> in complex with TAME showed densities (dark blue) for TAME (C-atoms in lime green) at both the IR tail-binding site (**c**), and the C-box binding site (**d**). Their positions overlap well with the crucial Arg78 and Tyr79 of Cdc20<sup>C<sup>box</sup></sup> and Ile498 and Arg499 of Cdc20<sup>IR</sup>, thereby inhibiting Cdc20 association (Extended Data Fig. 7).



APC/C (Extended Data Fig. 6e) is sufficient to compete for the auto-inhibitory segment at the C-box binding site.

Our findings suggest the interesting possibility of exploiting differences in the affinities of the two coactivators for the design of inhibitors that specifically target mitotic APC/C<sup>Cdc20</sup> and thus suppress cell proliferation. A small molecule, tosyl-L-arginine methyl ester (TAME), was reported to inhibit APC/C activation by both Cdc20 and Cdh1 through a proposed mechanism involving competition for the IR tail-binding site of coactivator<sup>29</sup>. However, we found that TAME is a more potent inhibitor of APC/C<sup>Cdc20</sup> than APC/C<sup>Cdh1</sup> (Fig. 4b). To understand the molecular basis underlying this inhibition, we determined the structure of APC/C<sup>ΔApc1-300s</sup> in complex with TAME (Extended Data Table 1). We observed TAME density not only at the IR tail-binding site, but also at the C-box binding site (Fig. 4c, d). This is consistent with the structural similarities of the IR tail and C-box bindings sites of Apc3A and Apc8B, respectively<sup>23</sup>, although there are critical differences. In Apc3A, TAME is reminiscent of the Ile–Arg motif of the coactivator IR tail, whereas in Apc8B, TAME resembles the Arg–Tyr/Phe of the coactivator C box (Extended Data Fig. 7). The mechanism of TAME inhibition of APC/C<sup>Cdc20</sup> through a tosyl-Arg motif to block the Cdc20<sup>Cbox</sup> binding site is reminiscent of the Apc1 auto-inhibitory segment that is displaced by APC/C phosphorylation. Both utilize a common structural motif (Arg–aromatic) that mimics Arg78 and Tyr79 of Cdc20<sup>Cbox</sup> to exploit the lower (relative to Cdh1) affinity of Cdc20 for the APC/C.

How phosphorylation regulates mitotic APC/C activation by Cdc20 has been a long-standing puzzle. Our *in vitro* studies show that of almost 150 phosphorylation sites in mitotic APC/C, only a few in Apc1 directly regulate Cdc20 binding through displacement of the auto-inhibitory segment. This work has relevance to understanding the control of other large multimeric complexes by multi-site phosphorylation.

**Online Content** Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

**Received 15 December 2015; accepted 6 April 2016.**

**Published online 27 April 2016.**

1. Pines, J. Cubism and the cell cycle: the many faces of the APC/C. *Nat. Rev. Mol. Cell Biol.* **12**, 427–438 (2011).
2. Primorac, I. & Musacchio, A. *Panta rhei*: the APC/C at steady state. *J. Cell Biol.* **201**, 177–189 (2013).
3. Chang, L. & Barford, D. Insights into the anaphase-promoting complex: a molecular machine that regulates mitosis. *Curr. Opin. Struct. Biol.* **29**, 1–9 (2014).
4. Kimata, Y., Baxter, J. E., Fry, A. M. & Yamano, H. A role for the Fizzy/Cdc20 family of proteins in activation of the APC/C distinct from substrate recruitment. *Mol. Cell* **32**, 576–583 (2008).
5. Van Voorhis, V. A. & Morgan, D. O. Activation of the APC/C ubiquitin ligase by enhanced E2 efficiency. *Curr. Biol.* **24**, 1556–1562 (2014).
6. Chang, L., Zhang, Z., Yang, J., McLaughlin, S. H. & Barford, D. Molecular architecture and mechanism of the anaphase-promoting complex. *Nature* **513**, 388–393 (2014).
7. Lahav-Baratz, S., Sudakin, V., Ruderman, J. V. & Hershko, A. Reversible phosphorylation controls the activity of cyclosome-associated cyclin-ubiquitin ligase. *Proc. Natl Acad. Sci. USA* **92**, 9303–9307 (1995).
8. Shteinberg, M., Protopopov, Y., Listovsky, T., Brandeis, M. & Hershko, A. Phosphorylation of the cyclosome is required for its stimulation by Fizzy/cdc20. *Biochem. Biophys. Res. Commun.* **260**, 193–198 (1999).
9. Kramer, E. R., Scheuringer, N., Podtelejnikov, A. V., Mann, M. & Peters, J. M. Mitotic regulation of the APC activator proteins CDC20 and CDH1. *Mol. Biol. Cell* **11**, 1555–1569 (2000).
10. Rudner, A. D. & Murray, A. W. Phosphorylation by Cdc28 activates the Cdc20-dependent activity of the anaphase-promoting complex. *J. Cell Biol.* **149**, 1377–1390 (2000).
11. Golan, A., Yudkovsky, Y. & Hershko, A. The cyclin-ubiquitin ligase activity of cyclosome/APC is jointly activated by protein kinases Cdk1-cyclin B and Plk. *J. Biol. Chem.* **277**, 15552–15557 (2002).
12. Kraft, C. et al. Mitotic regulation of the human anaphase-promoting complex by phosphorylation. *EMBO J.* **22**, 6598–6609 (2003).

13. Zachariae, W., Schwab, M., Nasmyth, K. & Seufert, W. Control of cyclin ubiquitination by CDK-regulated binding of Hct1 to the anaphase promoting complex. *Science* **282**, 1721–1724 (1998).
14. Jaspersen, S. L., Charles, J. F. & Morgan, D. O. Inhibitory phosphorylation of the APC regulator Hct1 is controlled by the kinase Cdc28 and the phosphatase Cdc14. *Curr. Biol.* **9**, 227–236 (1999).
15. Schwab, M., Neutznier, M., Möcker, D. & Seufert, W. Yeast Hct1 recognizes the mitotic cyclin Clb2 and other substrates of the ubiquitin ligase APC. *EMBO J.* **20**, 5165–5175 (2001).
16. Passmore, L. A. et al. Doc1 mediates the activity of the anaphase-promoting complex by contributing to substrate recognition. *EMBO J.* **22**, 786–796 (2003).
17. Vodermaier, H. C., Gieffers, C., Maurer-Stroh, S., Eisenhaber, F. & Peters, J. M. TPR subunits of the anaphase-promoting complex mediate binding to the activator protein CDH1. *Curr. Biol.* **13**, 1459–1468 (2003).
18. Patra, D. & Dunphy, W. G. Xp-p9, a *Xenopus* Suc1/Cks protein, is essential for the Cdc2-dependent phosphorylation of the anaphase-promoting complex at mitosis. *Genes Dev.* **12**, 2549–2559 (1998).
19. Shteinberg, M. & Hershko, A. Role of Suc1 in the activation of the cyclosome by protein kinase Cdk1/cyclin B. *Biochem. Biophys. Res. Commun.* **257**, 12–18 (1999).
20. Herzog, F., Mechtler, K. & Peters, J. M. Identification of cell cycle-dependent phosphorylation sites on the anaphase-promoting complex/cyclosome by mass spectrometry. *Methods Enzymol.* **398**, 231–245 (2005).
21. Steen, J. A. et al. Different phosphorylation states of the anaphase promoting complex in response to antimitotic drugs: a quantitative proteomic analysis. *Proc. Natl Acad. Sci. USA* **105**, 6069–6074 (2008).
22. Hegemann, B. et al. Systematic phosphorylation analysis of human mitotic protein complexes. *Sci. Signal.* **4**, rs12 (2011).
23. Chang, L., Zhang, Z., Yang, J., McLaughlin, S. H. & Barford, D. Atomic structure of the APC/C and its mechanism of protein ubiquitination. *Nature* **522**, 450–454 (2015).
24. Herzog, F. et al. Structure of the anaphase-promoting complex/cyclosome interacting with a mitotic checkpoint complex. *Science* **323**, 1477–1481 (2009).
25. Izawa, D. & Pines, J. Mad2 and the APC/C compete for the same site on Cdc20 to ensure proper chromosome segregation. *J. Cell Biol.* **199**, 27–37 (2012).
26. Yamaguchi, M. et al. Structure of an APC3-APC16 complex: insights into assembly of the anaphase-promoting complex/cyclosome. *J. Mol. Biol.* **427**, 1748–1764 (2015).
27. van Zon, W. et al. The APC/C recruits cyclin B1-Cdk1-Cks in prometaphase before D box recognition to control mitotic exit. *J. Cell Biol.* **190**, 587–602 (2010).
28. Sudakin, V., Shteinberg, M., Ganoth, D., Hershko, J. & Hershko, A. Binding of activated cyclosome to p13(suc1). Use for affinity purification. *J. Biol. Chem.* **272**, 18051–18059 (1997).
29. Zeng, X. et al. Pharmacologic inhibition of the anaphase-promoting complex induces a spindle checkpoint-dependent mitotic arrest in the absence of spindle damage. *Cancer Cell* **18**, 382–395 (2010).

**Supplementary Information** is available in the online version of the paper.

**Acknowledgements** This work was funded by the MRC Laboratory of Molecular Biology and a Cancer Research UK grant to D.B. PhD funding for S.Z. was from the Gates Cambridge Scholarship and Boehringer Ingelheim Fonds. C.A. is an EMBO Fellow. We are grateful to members of the Barford group for discussion; S. Chen, C. Savva and G. McMullan for EM facilities; J. Grimmett and T. Darling for computing; K. Zhang for advice on data processing; G. Murshudov for help with REFMAC and S. Aibara for advice on cloning.

**Author Contributions** S.Z. cloned the substrates and Cdh1 mutant, purified proteins and performed biochemical analysis. S.Z. and L.C. prepared grids, collected and analysed EM data and determined the three-dimensional reconstructions. S.Z. fitted coordinates, built models and made figures with help of L.C. C.A. cloned kinases and Cdc20 and established *in vitro* phosphorylation of the APC/C. Z.Z. and J.Y. cloned the APC/C mutants and the chimaeric proteins and prepared viruses. S.M. and M.S. performed mass spectrometry. D.B. directed the project and designed experiments with S.Z. S.Z. and D.B. wrote the manuscript with input from authors.

**Author Information** EM maps are deposited in the Electron Microscopy Data Bank with accession codes 3385 (APC/C<sup>Cdc20-Hs1</sup>), 3386 (apo unphosphorylated APC/C), 3387 (apo phosphorylated APC/C), 3388 (combined apo phosphorylated APC/C), 3389 (APC/C<sup>ΔApc1-300s</sup>) and 3390 (APC/C<sup>ΔApc1-300s-TAME</sup>). Protein coordinates are deposited in the Protein Data Bank under accession codes 5G04 (APC/C<sup>Cdc20-Hs1</sup>) and 5G05 (apo phosphorylated APC/C). Reprints and permissions information is available at [www.nature.com/reprints](http://www.nature.com/reprints). The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to D.B. ([dbarford@mrc-lmb.cam.ac.uk](mailto:dbarford@mrc-lmb.cam.ac.uk)).

## METHODS

No statistical methods were used to predetermine sample size. The experiments were not randomized and the investigators were not blinded to allocation during experiments and outcome assessment.

**Expression and purification of recombinant human APC/C.** The genes for recombinant human APC/C were cloned into a modified MultiBac system, expressed and purified as described<sup>30</sup>. The C terminus of Apc4 was fused to a TEV (tobacco etch virus)-cleavable StrepII $\times$ 2 tag.

**Protein kinase purification.** All four proteins (Cdk2, cyclin A3 (residues 174–432), Cks2 and Plk1 (residues 37–338)) for the kinases were expressed individually in BL21 (DE3) Star cells at 18 °C overnight. Pellets containing GST-tagged Cdk2, His-tagged cyclin A3 and His-SUMO-tagged Cks2 were combined and resuspended in the CDK lysis buffer (50 mM Tris/HCl pH 7.4, 180 mM NaCl, 5% glycerol and 2 mM DTT) supplemented with 0.1 mM PMSF, lysozyme, 5 units per ml benzonase and Complete EDTA-free protease inhibitors (Roche). After sonication, the cells were centrifuged at 20,000 r.p.m. for 1 h at 4 °C and the supernatant was incubated with the Glutathione Sepharose 4B (GE Healthcare) for 3 h at 4 °C. The resins were washed with the CDK lysis buffer and the GST-tag of Cdk2 was cleaved off with 3C PreScission protease overnight at 4 °C. The flow-through from the resins was collected and TEV-cleaved overnight at 4 °C. Finally, the protein complex was purified by a Superdex200 16/60 column (GE Healthcare).

The Polo kinase Plk1 with an N-terminal His-MBP tag was purified with a HisTrap HP column (GE Healthcare) in the PLK lysis buffer (50 mM Tris/HCl pH 7.5, 300 mM NaCl, 20 mM imidazole, 5% glycerol and 2 mM  $\beta$ -mercaptoethanol). The column was washed intensively with high-salt buffer (50 mM Tris/HCl pH 7.5, 1 M NaCl, 20 mM imidazole, 5% glycerol and 2 mM  $\beta$ -mercaptoethanol). Proteins were eluted with a gradient of the elution buffer (50 mM Tris/HCl pH 7.5, 300 mM NaCl, 300 mM imidazole, 5% glycerol and 2 mM  $\beta$ -mercaptoethanol) followed by TEV-cleavage overnight at 4 °C. The sample was re-applied onto the HisTrap HP column to remove the His tag, uncleaved proteins and nickel contaminations. The collected flow-through was concentrated and loaded onto a Superdex75 16/60 column (GE Healthcare) for final purification.

**In vitro phosphorylation of recombinant human APC/C.** Concentrated APC/C after Resource Q (GE Healthcare) was treated with Cdk2–cyclin A3–Cks2 and Plk1 in a molar ratio of 1:1.5 (APC/C: kinases) in a reaction buffer of 40 mM HEPES pH 8.0, 10 mM MgCl<sub>2</sub> and 0.6 mM DTT with 5 mM ATP and 50 mM NaF. The reaction mixture was incubated at 30 °C for 30 min before the final purification step by a Superose 6 3.2/300 column (GE Healthcare) in the APC/C gel-filtration buffer (20 mM HEPES pH 8.0, 150 mM NaCl, 0.5 mM TCEP).

**Expression and purification of the substrate Cdk2–cyclin A2–Cks2 and Cdc20.** Full-length human cyclin A2 was cloned into the pETM41 vector with an N-terminal His-MBP tag. The protein was expressed in BL21 (DE3) Star cells at 18 °C overnight. Pellets containing Cdk2, cyclin A2 and Cks2 were co-lysed for purification following a similar protocol as for the kinase purification.

Full-length human Cdc20 was cloned into a modified pFastBac HTa vector with an N-terminal His-MBP tag. The generated virus was amplified and expressed in Sf9 cells. Harvested cell pellets were resuspended in Cdc20 lysis buffer (50 mM HEPES pH 7.8, 500 mM NaCl, 30 mM imidazole, 10% glycerol and 0.5 mM TCEP) supplemented with 0.1 mM PMSF, 5 units per ml benzonase and Complete EDTA-free protease inhibitors and loaded onto a HisTrap HP column (GE Healthcare). Proteins were eluted with a gradient to 300 mM imidazole. Collected peak fractions were TEV-cleaved overnight in the dialysis bag (cut-off 6–8 kDa) against the dialysis buffer (50 mM HEPES, pH 7.8, 300 mM NaCl, 5% glycerol and 0.5 mM TCEP) at 4 °C. The protein was re-applied onto the HisTrap HP column and the flow-through was collected.

**Complex formation of APC/C<sup>Cdc20-Hsl1</sup> and APC/C <sup>$\Delta$ Apc1-300s-TAME</sup>.** *In vitro* phosphorylated APC/C was treated with 40  $\mu$ M CDK1/2 inhibitor III (ENZO Life Sciences) before incubating with purified Cdc20 and Hsl1 (with a molar ratio of 1:1.5:2) on ice. The complex was purified by a Superose 6 3.2/300 column using the Microakta system.

TAME (Sigma-Aldrich) was dissolved in 50% DMSO and 50% APC/C gel-filtration buffer at a concentration of 1 M. 4 mM TAME was added to purified APC/C <sup>$\Delta$ Apc1-300s</sup> and incubated on ice for 1 h before freezing cryo-grids.

**Ubiquitination assays.** The ubiquitination assay was performed with 60 nM recombinant human APC/C, 90 nM UBA1, 300 nM UbCH10, 300 nM Ube2S, 70  $\mu$ M ubiquitin, 2  $\mu$ M substrate Cdk2–cyclin A2–Cks2 or Hsl1, 5 mM ATP, 0.25 mg ml<sup>-1</sup> BSA, 15  $\mu$ M CDK1/2 inhibitor III and different concentrations of purified human Cdc20 (5–30 nM) or Cdh1 (5–30 nM) in a 10  $\mu$ l reaction volume with 40 mM HEPES pH 8.0, 10 mM MgCl<sub>2</sub> and 0.6 mM DTT (figure legends indicate the exact coactivator concentration used in each assay). Reaction mixtures were incubated at room temperature for various time points and terminated by

adding SDS/PAGE loading dye. Reactions were analysed by 4–12% NuPAGE Bis-Tris gels followed by western blotting with an antibody against the His-tag of ubiquitin (Clontech cat. code: 631212).

For the chimaeric Cdh1–Cdc20 assay, the following domain boundaries were used to generate chimaeric Cdh1–Cdc20 proteins: Cdh1<sup>NTD</sup> (residues 1–168), Cdh1<sup>WD40</sup> (residues 169–475), Cdh1<sup>IR</sup> (residues 476–496), Cdc20<sup>NTD</sup> (residues 1–165), Cdc20<sup>WD40</sup> (residues 166–475), Cdc20<sup>IR</sup> (residues 476–499). 10 nM of the chimaeric proteins were used in the assay.

**Peptide and TAME assays.** Eight peptides (Designer BioScience) spanning the Apc1 300s loop were synthesized for identification of the Apc1 auto-inhibitory segment. Each peptide contains 20 amino acids, with a ten-residue overlap with the neighbouring peptides: peptide 1 (LTAHLRLSKGDSPTSPFQ); peptide 2 (GDSPTSPFQNYSSHSQSR); peptide 3 (NYSSHSQSRSTSSPSLHRSR); peptide 4 (STSSPSLHRSRSPSISNMAAL); peptide 5 (SPSISNMAALSRASHPALGV); peptide 6 (SRASHPALGVHSFSGVQRFN); peptide 7 (HSFSGVQRFNISHNQSPKR) and peptide 8 (ISSHNQSPKRHSISHSPNSN). The following mutant peptides of peptide 7 were used to assess the relief of auto-inhibition: peptide R368E (HSFSGVQRFNISHNQSPKR), peptide 4S/E (HSFSGVQRFNIEEHNQEPKR) and peptide pS377 (HSFSGVQRFNISHNQphosphoSPKR). The peptides were dissolved at a concentration of 10 mM in 100% dimethylsulfoxide (DMSO) and diluted using the APC/C gel-filtration buffer. The ubiquitination assay was performed using a final concentration of 200  $\mu$ M peptide. The TAME assay was performed at a similar condition using 0.5–3 mM TAME and the substrate Hsl1.

**Size-exclusion chromatography to assess coactivator binding.** Purified APC/C samples (1 mg ml<sup>-1</sup>) were incubated with either Cdh1 or Cdc20 at a molar ratio of 1:1.5 on ice for 30 min. The sample was spun down at 13,000 r.p.m. for 5 min to remove any precipitates or aggregates before injecting onto a Superose 6 3.2/300 column using the Microakta system. The eluted peak fractions were analysed by SDS–PAGE and western blotting.

**Mass spectrometry.** Purified proteins were prepared for mass spectrometric analysis by in solution enzymatic digestion, without prior reduction and alkylation. Protein samples were digested with trypsin or elastase (Promega), both at an enzyme to protein ratio of 1:20. The resulting peptides were analysed by nano-scale capillary LC-MS/MS using an Ultimate U3000 HPLC (ThermoScientific Dionex) to deliver a flow of approximately 300 nl min<sup>-1</sup>. A C18 Acclaim PepMap100 5  $\mu$ m, 100  $\mu$ m  $\times$  20 mM nanoViper (ThermoScientific Dionex), trapped the peptides before separation on a C18 Acclaim PepMap100 3  $\mu$ m, 75  $\mu$ m  $\times$  250 mM nanoViper (ThermoScientific Dionex, San Jose, USA). Peptides were eluted with a 90 min gradient of acetonitrile (2% to 50%). The analytical column outlet was directly interfaced via a nano-flow electrospray ionization source, with a hybrid quadrupole orbitrap mass spectrometer (Q-Exactive Plus Orbitrap, ThermoScientific). LC-MS/MS data were then searched against an in-house LMB database using the Mascot search engine (Matrix Science)<sup>31</sup>, and the peptide identifications validated using the Scaffold program (Proteome Software Inc.)<sup>32</sup>. All data were additionally interrogated manually.

**Electron microscopy.** Freshly purified APC/C samples were first visualized by negative-staining EM to check the sample quality and homogeneity and to get initial low-resolution reconstructions. Micrographs were recorded on an FEI Spirit electron microscope at an accelerating voltage of 120 kV and at a defocus of approximately –1.5  $\mu$ m. For cryo-EM, 2- $\mu$ l aliquots of the sample at ~0.15 mg ml<sup>-1</sup> were applied onto the Quantifoil R2/2 grids coated with a layer of continuous carbon film (approximately 50 Å thick). Grids were treated with a 9:1 argon:oxygen plasma cleaner for 20 to 40 s before use. The grids were incubated for 30 s at 4 °C and 100% humidity before blotting for 5 s and plunging into liquid ethane using an FEI Vitrobot III. The grids were loaded into an FEI Tecnai Polara electron microscope at an acceleration voltage of 300 kV. Micrographs were taken using EPU software (FEI) at a nominal magnification of 78,000 which yields a pixel size of 1.36 Å per pixel. They were recorded by an FEI Falcon III direct electron detector with a defocus range of –2.0 to –4.0  $\mu$ m. The exposure time for each micrograph was 2 s at a dose rate of 27 electrons per Å<sup>2</sup> per s. 34 movie frames were recorded for each micrograph as described<sup>33</sup>.

**Image processing.** All movie frames were aligned by the motioncorr program<sup>34</sup> before subsequent processing. First, the contrast transfer function parameters were calculated with CTFFIND3 or Gctf<sup>35,36</sup>. Particles in 264 pixels  $\times$  264 pixels were selected by automatic particle picking in RELION 1.4 (ref. 37). The following steps were performed to exclude bad particles from the data set: (1) automatically picked particles in each micrograph were screened manually to remove ice contaminations<sup>38</sup>; (2) after particle sorting in RELION, particles with poor similarity to reference images were deleted; (3) 2-dimensional classification was performed and particles in bad classes with poorly recognizable features were excluded. The remaining particles were divided into six classes using three-dimensional



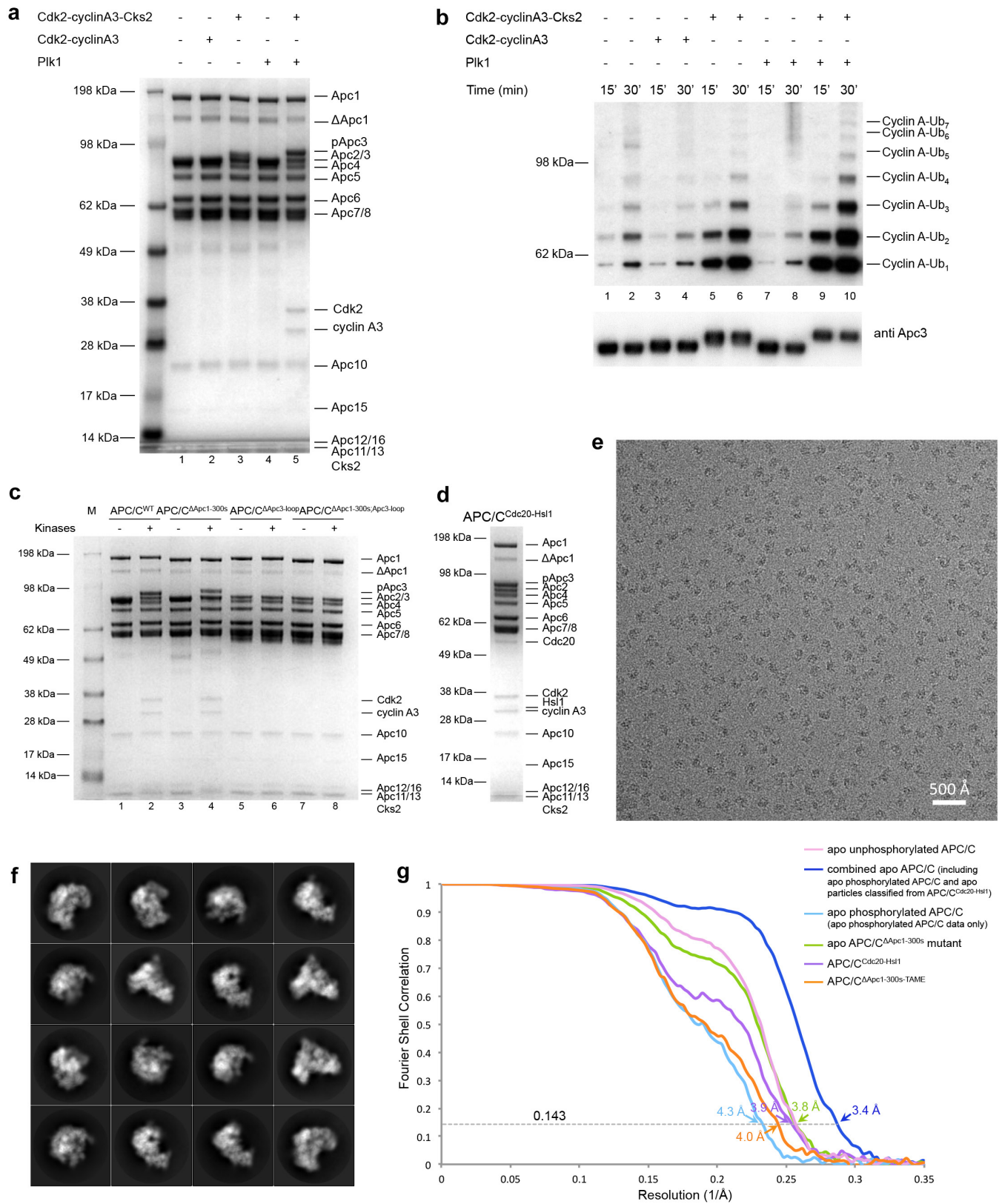
classification in RELION. During this process particles with conformational heterogeneity and leftover bad particles were removed from the final reconstruction. After 3D refinement, beam-induced particle motion was corrected using particle polishing in RELION<sup>33,39</sup>. All resolution estimations were based on the gold-standard Fourier shell correlation (FSC) calculations using the FSC = 0.143 criterion. The model for the apo state (Fig. 2a, b, combined apo APC/C structure) was built based on a 3.4 Å resolution map, reconstructed by combined data collected from pure apo phosphorylated APC/C and the apo state particles classified from the APC/C<sup>Cdc20-Hsl1</sup> complex (Extended Data Fig. 2 and Extended Data Table 1). This map allowed for model building of regions which were less well-resolved and presented as poly-alanines in the previous APC/C<sup>Cdh1-Emi1</sup> structure<sup>23</sup>, including Apc2<sup>NTD</sup>, Apc1 loops in both the WD40 domain and the middle domain and regions in Apc5<sup>NTD</sup>. A summary of all EM reconstructions obtained in this work is listed in Extended Data Table 1a.

**Model building.** Model building of both apo APC/C and APC/C<sup>Cdc20-Hsl1</sup> structures were performed in COOT<sup>40</sup>. Initially, available atomic structure of human APC/C<sup>Cdh1-Emi1</sup> (PDB 4UI9)<sup>23</sup> and the crystal structure of Cdc20<sup>WD40</sup> (PDB 4GGC)<sup>41</sup> were rigid-body fitted in individual subunits into the cryo-EM maps in Chimera<sup>42</sup>. All fitted structures were rebuilt according to the cryo-EM map. Cdc20<sup>NTD</sup>, Cdc20<sup>CTD</sup> and several loop regions not seen in previous structures were built *ab initio*. The models were refined by REFMAC 5.8 (ref. 43). A REFMAC weight of 0.04 was defined by cross-validation using half reconstructions<sup>44</sup>. A resolution limit of 4.0 Å or 3.5 Å was used for the APC/C<sup>Cdc20-Hsl1</sup> and the apo APC/C structure, respectively. All available crystal structures or NMR structures were used for secondary structure restraints. The refinement statistics are summarized in Extended Data Table 1b.

**Map visualization.** Figures were generated using Pymol and Chimera<sup>42</sup>.

**Sequence alignment.** Sequence alignment was performed using Jalview<sup>45</sup>.

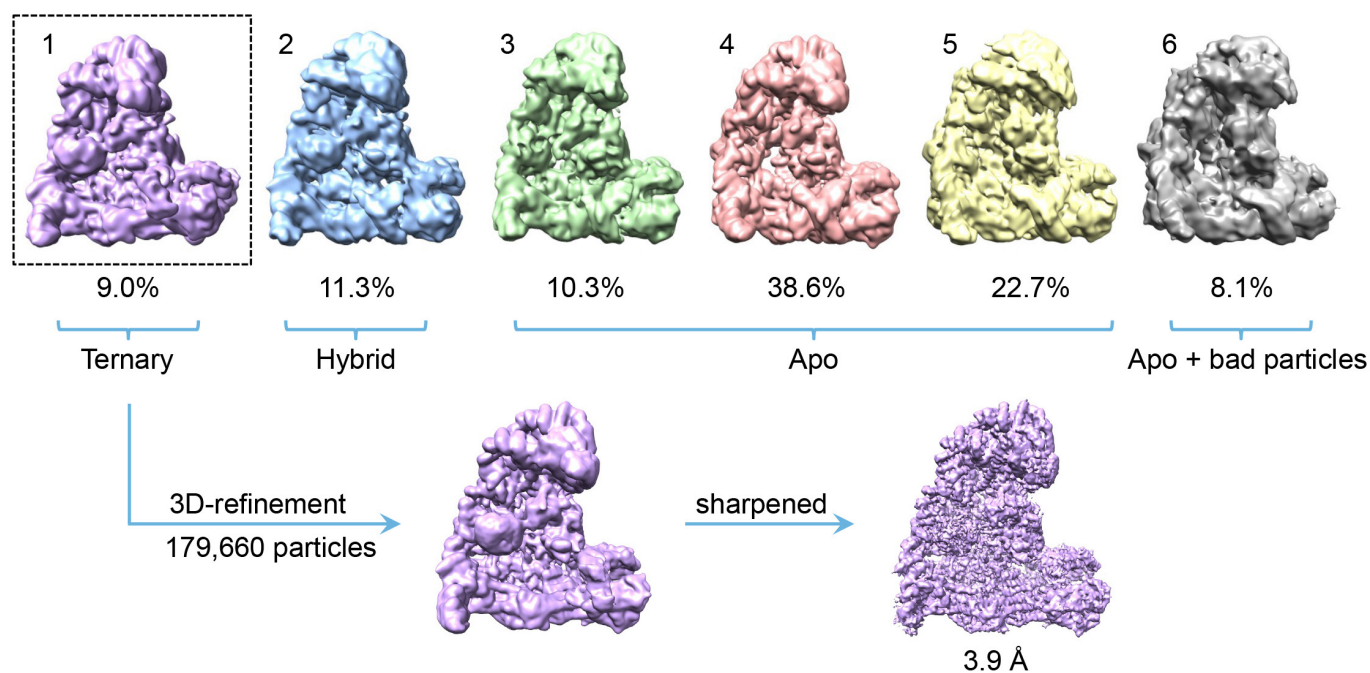
30. Zhang, Z. *et al.* Recombinant expression, reconstitution and structure of human anaphase-promoting complex (APC/C). *Biochem. J.* **449**, 365–371 (2013).
31. Perkins, D. N., Pappin, D. J., Creasy, D. M. & Cottrell, J. S. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* **20**, 3551–3567 (1999).
32. Keller, A., Nesvizhskii, A. I., Kolker, E. & Aebersold, R. Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. *Anal. Chem.* **74**, 5383–5392 (2002).
33. Bai, X. C., Fernandez, I. S., McMullan, G. & Scheres, S. H. Ribosome structures to near-atomic resolution from thirty thousand cryo-EM particles. *eLife* **2**, e00461 (2013).
34. Li, X. *et al.* Electron counting and beam-induced motion correction enable near-atomic-resolution single-particle cryo-EM. *Nat. Methods* **10**, 584–590 (2013).
35. Mindell, J. A. & Grigorieff, N. Accurate determination of local defocus and specimen tilt in electron microscopy. *J. Struct. Biol.* **142**, 334–347 (2003).
36. Zhang, K. Gctf: real-time CTF determination and correction. *J. Struct. Biol.* **193**, 1–12 (2016).
37. Scheres, S. H. RELION: implementation of a Bayesian approach to cryo-EM structure determination. *J. Struct. Biol.* **180**, 519–530 (2012).
38. Scheres, S. H. Semi-automated selection of cryo-EM particles in RELION-1.3. *J. Struct. Biol.* **189**, 114–122 (2015).
39. Scheres, S. H. Beam-induced motion correction for sub-megadalton cryo-EM particles. *eLife* **3**, e03665 (2014).
40. Emsley, P. & Cowtan, K. Coot: model-building tools for molecular graphics. *Acta Crystallogr. D* **60**, 2126–2132 (2004).
41. Tian, W. *et al.* Structural analysis of human Cdc20 supports multisite degron recognition by APC/C. *Proc. Natl Acad. Sci. USA* **109**, 18419–18424 (2012).
42. Yang, Z. *et al.* UCSF Chimera, MODELLER, and IMP: an integrated modeling system. *J. Struct. Biol.* **179**, 269–278 (2012).
43. Murshudov, G. N. *et al.* REFMAC5 for the refinement of macromolecular crystal structures. *Acta Crystallogr. D* **67**, 355–367 (2011).
44. Fernández, I. S., Bai, X. C., Murshudov, G., Scheres, S. H. & Ramakrishnan, V. Initiation of translation by cricket paralysis virus IRES requires its translocation in the ribosome. *Cell* **157**, 823–831 (2014).
45. Waterhouse, A. M., Procter, J. B., Martin, D. M., Clamp, M. & Barton, G. J. Jalview Version 2—a multiple sequence alignment editor and analysis workbench. *Bioinformatics* **25**, 1189–1191 (2009).



Extended Data Figure 1 | See next page for caption.

**Extended Data Figure 1 | Preparations and EM images of different APC/C samples used for structural studies.** **a**, Recombinant human APC/C was phosphorylated *in vitro* using Cdk2–cyclin A3, Cdk2–cyclin A3–Cks2 or Plk1 alone or with both Cdk2–cyclin A3–Cks2 and Plk1. The phosphorylated APC/C samples are shown after SDS–PAGE. **b**, *In vitro* phosphorylated recombinant human APC/C can be fully activated by Cdc20 to ubiquitylate a native substrate Cdk2–cyclin A2–Cks2 when both kinases were added (lanes 9, 10). Without Cks2 (lanes 3, 4) or with Plk1 alone (lanes 7, 8) no activation of the APC/C could be observed, whereas treating with Cdk2–cyclin A3–Cks2 alone (lanes 5, 6) resulted in its partial activation. Samples were recorded at 15 and 30 min of the reaction and 20 nM of Cdc20 was used. This experiment was replicated three times.

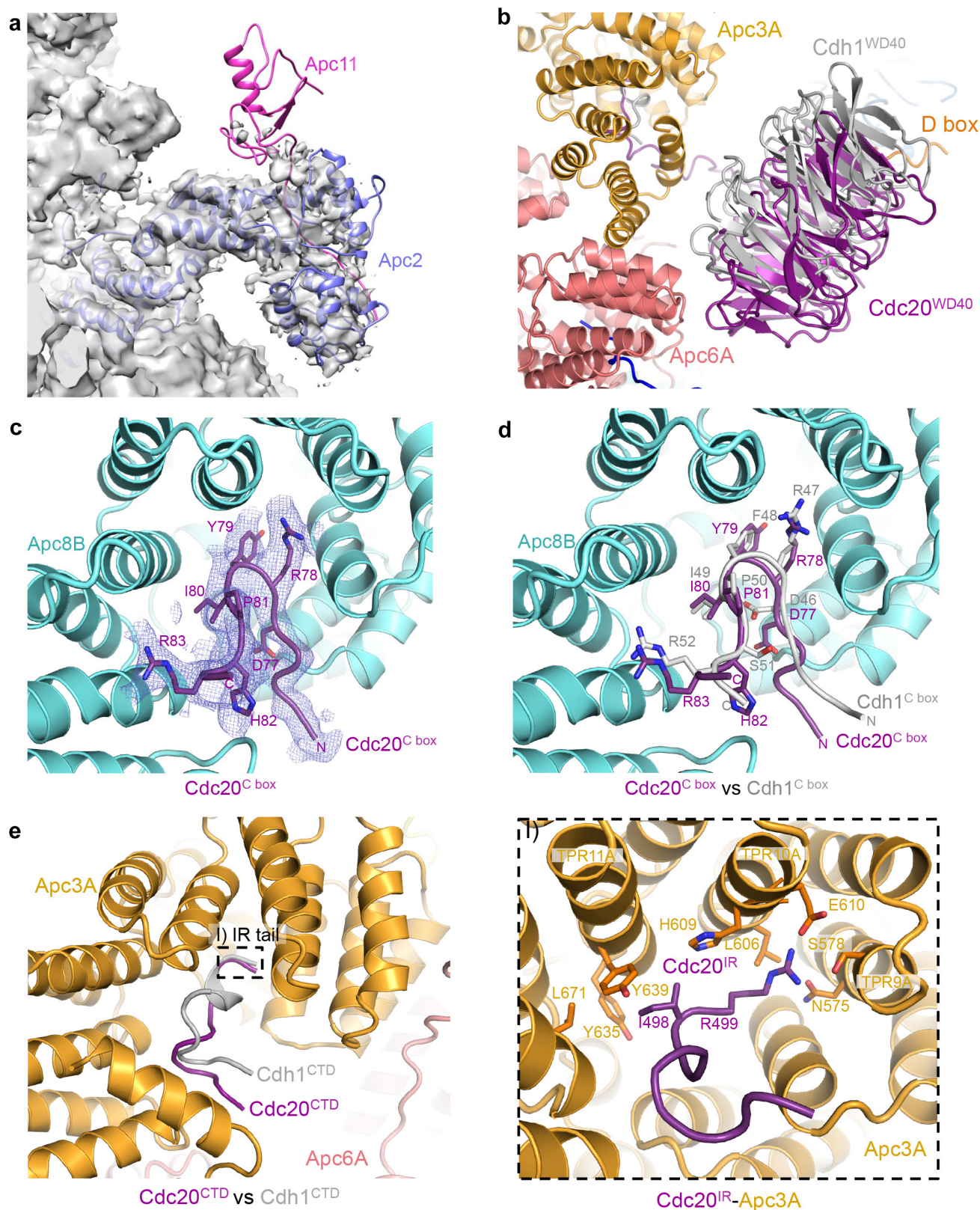
Anti-Apc3 antibodies (BD Bioscience, cat. code: 610454) were used as a loading control. **c**, Purified wild-type (WT) APC/C and mutant samples with and without kinase treatment (both Cdk2–cyclin A3–Cks2 and Plk1). Upon deletion of the Apc3 loop, no association of the Cdk2–cyclin A3–Cks2 kinase to the APC/C could be observed (lanes 6 and 8). **d**, SDS–PAGE of purified APC/C<sup>Cdc20-Hsl1</sup> ternary complex. **e**, A typical cryo-EM micrograph of APC/C<sup>Cdc20-Hsl1</sup> representative of 15,582 micrographs. **f**, Gallery of two-dimensional averages of APC/C<sup>Cdc20-Hsl1</sup> showing different views; representative of 100 two-dimensional averages. **g**, Gold-standard FSC curves of all APC/C reconstructions in this work. See Supplementary Fig. 1 for gel source data.



**Extended Data Figure 2 | Three-dimensional classification of APC/ $C^{Cdc20-Hsl}$ .** The initial particles after two-dimensional classification were divided into six classes by three-dimensional classification using RELION. The resultant classes were grouped into four categories: (i) 9.0% in the active ternary state with coactivator and substrate bound; (ii) 11.3% in a hybrid state with coactivator bound, but the APC/C in the inactive

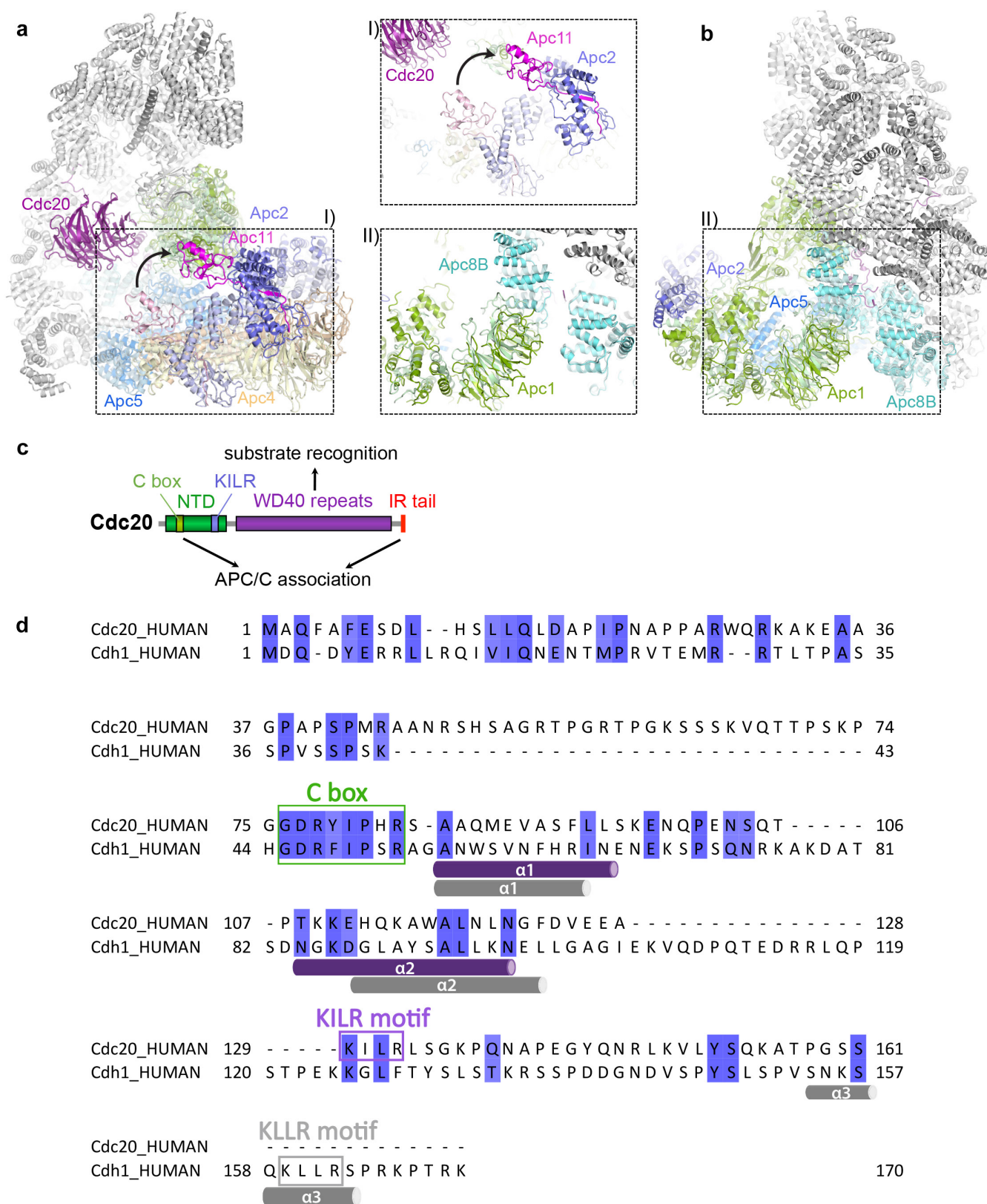
conformation; (iii) 71.6% in the inactive apo state; (iv) 8.1% with poorer reconstruction owing to some bad particles. The first class in the active ternary state containing 179,660 particles was used for three-dimensional refinement and movie correction to obtain the final reconstruction at 3.9 Å.





**Extended Data Figure 3 | Comparison of Cdc20 and Cdh1 association to the APC/C.** **a**, The catalytic module (Apc2-Apc11) of the APC/C<sup>Cdc20-Hsl1</sup> complex is flexible and almost no density accounting for Apc11 (pink, modelled based on the structure of APC/C<sup>Cdh1-Emi1</sup>, PDB 4UI9)<sup>23</sup> could be observed. **b**, The WD40 domain of Cdc20 (purple) occupies a similar position as Cdh1<sup>WD40</sup> (grey), but it is displaced from the APC/C by as

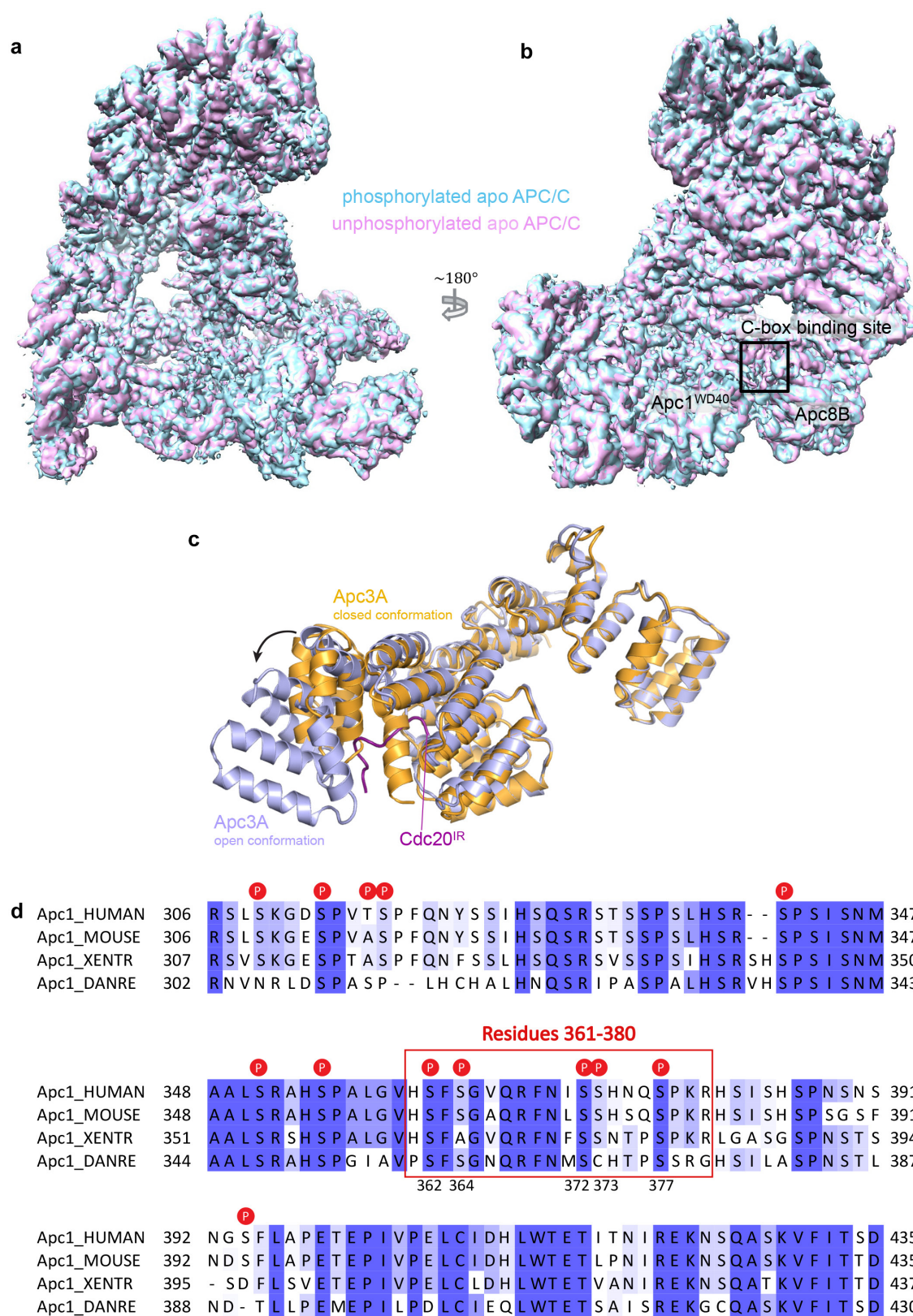
much as 10 Å. **c**, **d**, EM density for Cdc20<sup>C box</sup> allowed for *ab initio* model building and the C-box interaction with Apc8B (cyan) is well conserved between the two coactivators. **e**, Both Cdc20<sup>IR</sup> (right) and Cdh1<sup>IR</sup> (left) associates with Apc3A (orange), although the EM density for Cdc20<sup>IR</sup> is much weaker (not shown) and the C-terminal  $\alpha$ -helix in Cdh1<sup>IR</sup> is absent.



**Extended Data Figure 4 | Conformational changes of the APC/C between the inactive apo and the active ternary states and domain and sequence analysis of Cdc20.** **a, b,** Subunits that undergo conformational changes upon coactivator and substrate binding are highlighted in their ternary state and coloured as in Fig. 1, while the corresponding proteins in the inactive apo state are in lighter shades. In the active conformation, the

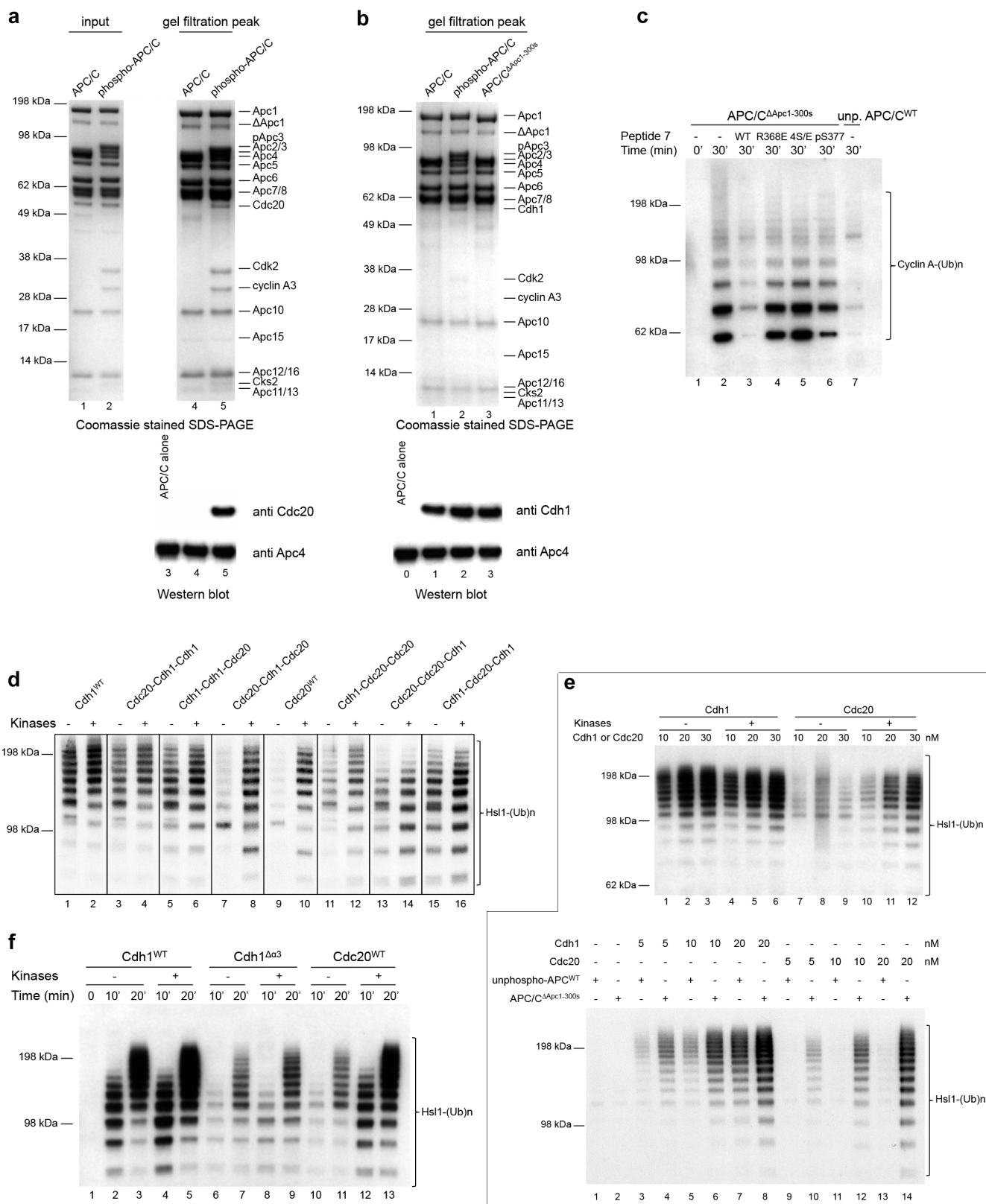
platform subdomain containing subunits Apc1, Apc4 and Apc5 is shifted upward, inducing a large movement of the catalytic module to enable E2 access. **c,** Domain organization of Cdc20. **d,** Sequence alignment of Cdc20<sup>NTD</sup> and Cdh1<sup>NTD</sup> with  $\alpha$ -helices represented as cylinders (purple and grey for Cdc20<sup>NTD</sup> and Cdh1<sup>NTD</sup>, respectively) underneath the sequences and the C-box and KILR/KLLR motif highlighted.





**Extended Data Figure 5 | Comparison of apo APC/C in unphosphorylated and phosphorylated states. a, b,** Superposition of the apo unphosphorylated (magenta) and phosphorylated (cyan) APC/C EM maps revealed little conformational differences except in the vicinity of the C-box binding site. **c,** Apc3A is in an equilibrium between open (light blue) and closed (orange) conformations. While in the inactive apo

state, the majority of Apc3A is in the closed state, association of Cdc20<sup>IR</sup> stabilizes the open state. **d,** Sequence alignment of the Apc1 300s loop across different species human, mouse, *Xenopus tropicalis* (western clawed frog) and *Danio rerio* (zebrafish). Phosphorylation sites are indicated and residues accounting for the Apc1 auto-inhibitory segment (361–380) are boxed.



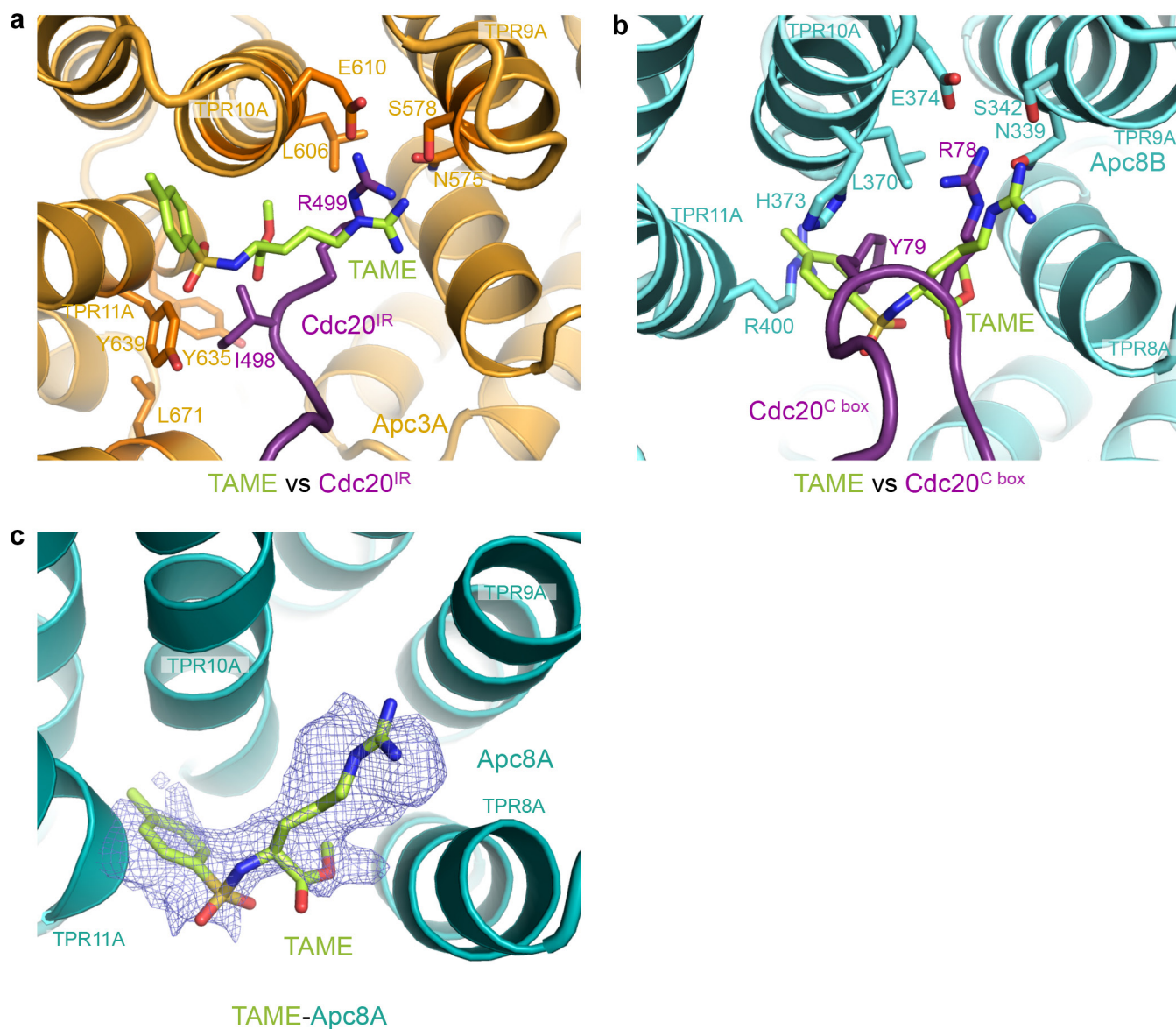
Extended Data Figure 6 | See next page for caption.



**Extended Data Figure 6 | Analytical gel filtration and activity assays.**

**a**, With equal amount of input Cdc20, phosphorylated APC/C could form a stable binary complex with Cdc20 after a gel-filtration purification step (lane 5), whereas unphosphorylated APC/C could not (lane 4). **b**, Both unphosphorylated and phosphorylated APC/C associate with Cdh1 stably on gel filtration, as well as APC/C<sup>ΔApc1-300s</sup>. Anti Cdc20 antibody (Santa Cruz Biotechnology, cat. code: sc-8358) and anti Cdh1 antibody (Sigma, cat. code: C7855) were used for detection; antibody to Apc4 (ref. 6) served as a loading control and unphosphorylated APC/C alone is used as a negative control for western blotting. **c**, Point mutations of peptide 7 (residues 361–380), either when Arg368 was mutated to glutamate or when the four neighbouring serines were mutated to phospho-mimics (Glu), caused the peptide to abolish its inhibition effect and restored the APC/C activity (lanes 4, 5). Phosphorylation of a single Ser377 only resulted in partial activation of the APC/C (lane 6). **d**, Chimaeric proteins composed of the NTD, the WD40 domain and the IR tail of either Cdc20 or Cdh1 were purified to study their differences

in APC/C activation. Both the NTD and the CTD of the coactivators are essential for their association with the APC/C. Swapping both NTD and CTD of Cdh1 with Cdc20 makes it phosphorylation sensitive (lanes 7, 8), similar to Cdc20 (lanes 9, 10) and vice versa. **e**, Top, Cdh1 can activate both unphosphorylated and phosphorylated APC/C similarly, whereas Cdc20 requires APC/C phosphorylation for its activity. Bottom, a titration of Cdh1 against unphosphorylated APC/C and APC/C<sup>ΔApc1-300s</sup> showed enhanced activity in the absence of the Apc1 auto-inhibitory segment at low Cdh1 concentration ( $\leq 10$  nM), whereas Cdc20 requires displacement of the auto-inhibitory segment for its activity. **f**, Deletion of the Cdh1  $\alpha$  3 helix resulted in reduced activation of the APC/C and makes Cdh1 more phosphorylation sensitive. The substrate Cdk2–cyclin A2–Cks2 was used for assay in **c** and Hsl1 for the assays in **d–f**. 20 nM Cdc20 was used in **c**, 10 nM chimaeric coactivators in **d** and 30 nM coactivators in **f**. Experiments in **a** and **b** were replicated two times, in **c**, **e** and **f** three times and in **d** four times. See Supplementary Fig. 1 for gel source data.



**Extended Data Figure 7 | TAME competes with Cdc20 to bind at the IR-tail and the C-box binding sites. a,** TAME (C atoms in lime green) is superimposed with Cdc20<sup>IR</sup> (purple) and the arginine motif in both structures engages the same binding site on Apc3A (orange). **b,** The

tosyl-Arg motif of TAME overlaps with Arg78–Tyr79 of Cdc20<sup>C box</sup> at the C-box binding site to out-compete Cdc20. **c,** A density for TAME was also observed within a pocket of the Apc8A TPR super-helix, similar to that of Apc8B.

Extended Data Table 1 | EM data collection, processing statistics and structure refinement statistics

**a. Statistics of all cryo-EM reconstructions**

Samples	Particles used for final reconstruction	Resolution (Å)	EM-DB accession code
APC/C <sup>Cdc20-Hsl1</sup>	179,660 (9.0%)	3.9	3385
apo unphosphorylated APC/C	347,317 (49.5%)	3.8	3386
apo phosphorylated APC/C (apo phosphorylated APC/C data only)	83,642 (63.1%)	4.3	3387
combined apo phosphorylated APC/C (both apo phosphorylated APC/C data and apo phosphorylated particles classified from APC/C <sup>Cdc20-Hsl1</sup> )	921,993 (70.8%)	3.4	3388
apo APC/C <sup>ΔApc1-300s</sup> mutant	262,090 (80.2%)	3.8	3389
APC/C <sup>ΔApc1-300s-TAME</sup>	246,065 (78.2%)	4.0	3390

**b. Statistics of APC/C<sup>Cdc20-Hsl1</sup> and combined apo phosphorylated APC/C structure determination**

<b>Data collection</b>	APC/C <sup>Cdc20-Hsl1</sup>	combined apo phospho. APC/C
EM	FEI Polara, 300k eV	FEI Polara, 300k eV
Detector	FEI Falcon III	FEI Falcon III
Pixel size (Å)	1.36	1.36
Defocus range (μm)	2.0-4.0	2.0-4.0
<b>Reconstruction</b>		
Software	RELION 1.4	RELION 1.4
Accuracy of rotations (degrees)	1.268	0.884
Accuracy of translations (pixels)	0.81	0.62
Final resolution (Å)	3.9	3.4
<b>Refinement</b>		
Software	RefMac 5.8	RefMac 5.8
Refmac weight	0.04	0.04
Resolution limit (Å)	4.0	3.5
Residue number	8164	7908
Average Fourier shell correlation	0.7568	0.7778
R factor	0.3621	0.3451
Rms bond length (Å)	0.0121	0.0141
Rms bond angle (°)	1.7812	1.7435
<b>Validation</b>		
Ramachandran plot		
Preferred	7502 (91.89%)	7427 (93.92%)
Allowed	400 (4.90%)	287 (3.63%)
Outliers	262 (3.21%)	194 (2.45%)
<b>RCSB PDB accession code</b>	5G04	5G05

**Extended Data Table 2 | Phosphorylation sites of Apc1 and Apc3 subunits of *in vitro* phosphorylated APC/C identified by mass spectrometry**

Protein	Phospho-sites	APC/C <sup>WT</sup>	APC/C <sup>WT</sup>	APC/C <sup>ΔApc3-loop</sup>	APC/C <sup>WT</sup>	APC/C <sup>WT</sup>	APC/C <sup>WT</sup>	Comparison with published studies Ref.
Kinases used	No treatment	Cdk2-cyclinA3	Cdk2-cyclinA3-Cks2 + Plk1	Cdk2-cyclinA3-Cks2	Plk1	Cdk2-cyclinA3-Cks2 + Plk1		
Apc1	LVG560LQE							22
	QEVY059HE							22(interphase)
	PPG5202FRE							12,21,22
	LFSG233SRV	N.D.				N.D.		21
	LKF5286EQG							12,22
	QGGT291PQN							12,21,22
	NVAT297SSS							
	VAT5298SSL							
	ATSS399SLT							
	RSL5309KGD			N.D.				22(interphase)
	KGD5313PVT							22
	SPVT316SPF							22
	PVT5317PFQ							22
	HSR5341PSI							12,21,22
	RSP5343GN							12,22
	AAL5351RAH							22
	RAH5355PAL							12,21,22
	GVHS362FSG							12,21,22
	HSF5364GVQ							12,22
	FNS5372SHN							22
	NIS5373HNQ							12,22
	HNQ5377PKR							12,21,22
	ISH5386PNS					N.D.		22(interphase)
	SPNS5389NSN					N.D.		22
	SNG5394FLA							
	WTET416ITN							
	VLYT501GVV							
	PAP5318LTM	N.D.	N.D.	N.D.				22
	PSLT520MSN	N.D.	N.D.					12(interphase),22
	LTM522NTM	N.D.	N.D.					22(interphase)
	MSMT524MPR	N.D.	N.D.					22(interphase)
	RPST530PLD							12(interphase),21,22
	DGV5336TPK							
	GVST537PKP							12,22
	KPL5342KLL		N.D.					21
	LLG5347LDE							12,22
	VLL5355PVP							12,21,22
	LRD5363SKL							22
	RDS5364KLH	N.D.						21,22
	LHD5369LYN							12(interphase),22
	EDCT576FQQ							
	QLG5382YH							
	LEL560NGS			N.D.				21
	FEH5686LSP							22
	GSL5688PVI							12,21,22
	ARP5698ETG							22
	PSET701GSD							22
	ETG5703DDD							22(interphase)
	LCL5731PSE							12,22(interphase),21
	NRF5916FRH							22
	FRH5920TSV							
	RHS5921SVS							
	HST5922VSS							
	VLS51001DVP							
	KHK51347PSY							
	KSP51349YQ							

The pink shading shows the presence of phosphorylation sites and the white indicates its absence.  
The phosphorylation sites within the Apc1 AI segment are highlighted in light purple shadows.

Protein	Phospho-sites	APC/C <sup>WT</sup>	APC/C <sup>WT</sup>	APC/C <sup>ΔApc3-loop</sup>	APC/C <sup>WT</sup>	APC/C <sup>WT</sup>	APC/C <sup>WT</sup>	Comparison with published studies Ref.
Kinases used	No treatment	Cdk2-cyclinA3	Cdk2-cyclinA3-Cks2 + Plk1	Cdk2-cyclinA3-Cks2	Plk1	Cdk2-cyclinA3-Cks2 + Plk1		
Apc3	LPN5183CTT							22
	NGC11857QV							
	SCTT186QVP							
	PNH5192LSH							
	HSL5194HRQ							
	QPET200VLT							
	TVLT203ETP							21
	LTE205PQD							12,21,22
	PQD2109EL							12
	NLE5219SNS	N.D.						22(interphase)
	LES5220NSK	N.D.						21,22
	SSNS222KYS							22
	SKYS225LNT							
	SLNT228DSS							
	NTD5230SVS							
	TSD5231VSF							21,22
	SSVS233YID							
	YID5237AVI							
	AVIS241PDT							22
	SPDT244VPL							12,22
	GTGT251SIL							
	TGT5252SIL							22
	SIL5255KQV							
	KPKT264GRS	N.D.				N.D.		21
	TGR5267LLG					N.D.		
	AAL5276PLT							22
	SPLT279PSF							22
	LTP5281FGI							
	PLET289PSF							
	ETP291PGD						N.D.	12
	QNYT302NTP							12,22
	YTN304PPV							12,22
	DVPS312TGA							12,22
	VPS313GAP							12,22
	RGCT327GK	N.D.	N.D.			N.D.		22
	QTGT329KSV							
	GTK5331VFS	N.D.	N.D.			N.D.		
	SVFS334QSG	N.D.				N.D.		21
	FSQS336GNS	N.D.				N.D.		21
	REV3343PIL							12
	LAQT349QSS							
	QTQS351SGP							
	TQS352GPQ							
	GPQT356STT							
	PQT3557TTP							
	QTS358TPQ							
	TSTT359PQV							
	QVL364PTI							
	LSP3366TS							12
	TTT3369PPN							12,21,22
	RLFT338SDS							21
	LFTS394DSS							21
	TSD3386STT							21
	SDS3387TTK							21
	DSST338TKE							
	SSTT338KEN							21
	GGT419QPN							
	IND5426LEI							12,21
	LEIT430KLD							12
	KLD5434SII							12
	LDS5435BS							12,21
	SII5438EOK							21
	GKS5443TIT							
	KIS5444ITP							21,22
	STIT5446PQI							12,21
	MNF5761WAM							
	IMG7800DES							
	TOE5803QES							
	QGE5806SHY							
	QES5807MTD							
	SSMT809DAD							
	ADD7814QLH							
	AAE5821DEF							



**Extended Data Table 3 | Summary of phosphorylation sites of *in vitro* phosphorylated APC/C subunits (excluding Apc1 and Apc3) identified by mass spectrometry**

Protein	Phospho-sites	APC/C <sup>WT</sup> No treatment	APC/C <sup>WT</sup> Cdk2-cyclinA3- Cks2 + Plk1	Comparison with published studies Ref.
Apc2	ELD <b>S205</b> RYA			
	LLQ <b>S218</b> PLC			21
	RPA <b>S314</b> PEA			12,21,22
	SLET <b>S466</b> GQD			
	GQD <b>S470</b> EDD			22
	EDD <b>S474</b> GEP			
	HQF <b>S532</b> FSP			21
	FSF <b>S534</b> PER			12,21,22
	LID <b>S732</b> DDE			22(interphase)
	DDE <b>S736</b> DSG			22(interphase)
	ESD <b>S738</b> GMA			22(interphase)
Apc4	GMA <b>S742</b> QAD			
	ARV <b>T199</b> GIA			
	KGK <b>Y469</b> FNV			12
	DLV <b>S488</b> PPN			
	LDE <b>S757</b> SDE			
	DES <b>S758</b> DEE			
Apc5	EVL <b>S777</b> ESE			21,22
	LSE <b>S779</b> EAE			12
	PMMT <b>T15</b> NGV			21
	HKT <b>S130</b> VVG			
	MEL <b>T178</b> SRD			22
	ELT <b>S179</b> RDE			22
	LDV <b>S195</b> VRE			12,21
	QQA <b>S221</b> LLK			21,22
	NDE <b>T228</b> KAL			
	KAL <b>T232</b> PAS			21
Apc6	VAS <b>S674</b> AAS			21
	KDE <b>S112</b> GFK			12,22
	NII <b>S559</b> PPW			12,21,22
	EKQ <b>T573</b> AEE			
	AEET <b>S77</b> GLT			
	TGL <b>T580</b> PLE			12,21,22
	PLE <b>T584</b> SRK			
	LET <b>S585</b> RKT			21,22
	TPD <b>S592</b> RPS		N.D.	
	SRP <b>S595</b> LEE		N.D.	12
	LEET <b>S599</b> FEI			12
	MNE <b>S607</b> DMM			
	LET <b>S614</b> MSD			
1	2	3	4	5

Protein	Phospho-sites	APC/C <sup>WT</sup> No treatment	APC/C <sup>WT</sup> Cdk2-cyclinA3- Cks2 + Plk1	Comparison with published studies Ref.
Apc7	VRP <b>S119</b> TGN			12
	RPST <b>T120</b> GNS			12
	TGN <b>S123</b> AST			12,22
	NSA <b>S125</b> TPQ			12,22
	SAS <b>T126</b> PQS			21,22
	MEG <b>S573</b> GEE			
	LEG <b>S582</b> DSE			
	GSD <b>S584</b> EAA			
Apc8	QGE <b>T562</b> PTT			12,21
	TPT <b>T565</b> EVP			12,21
	ANNT <b>S582</b> PTR			12,21,22
	NTP <b>T584</b> RRV			12
	RRV <b>S588</b> PLN			21,22
	LNL <b>S593</b> SVT			22
Apc12	SSV <b>T596</b> P			21,22
	VGG <b>S42</b> DGE			
	IGL <b>S51</b> SDP			
	GLS <b>S52</b> DPK			
Apc15	DPK <b>S56</b> REQ	N.D.		
	NRS <b>S78</b> QFG			
Apc16	DED <b>S76</b> EED			
	EED <b>S80</b> EDD			
	YNE <b>S98</b> PDD			
Apc16	SSS <b>S8</b> SAG	N.D.		
	VSG <b>S16</b> SVT	N.D.		
	FSV <b>S26</b> DLA	N.D.		
1	2	3	4	5

The pink shading shows the presence of phosphorylation sites and the white indicates its absence.

# Activation of the A<sub>2A</sub> adenosine G-protein-coupled receptor by conformational selection

Libin Ye<sup>1,2</sup>, Ned Van Eps<sup>2</sup>, Marco Zimmer<sup>2,3</sup>, Oliver P. Ernst<sup>2,4</sup> & R. Scott Prosser<sup>1,2</sup>

Conformational selection and induced fit are two prevailing mechanisms<sup>1,2</sup> to explain the molecular basis for ligand-based activation of receptors. G-protein-coupled receptors are the largest class of cell surface receptors and are important drug targets. A molecular understanding of their activation mechanism is critical for drug discovery and design. However, direct evidence that addresses how agonist binding leads to the formation of an active receptor state is scarce<sup>3</sup>. Here we use <sup>19</sup>F nuclear magnetic resonance to quantify the conformational landscape occupied by the adenosine A<sub>2A</sub> receptor (A<sub>2A</sub>R), a prototypical class A G-protein-coupled receptor. We find an ensemble of four states in equilibrium: (1) two inactive states in millisecond exchange, consistent with a formed (state S<sub>1</sub>) and a broken (state S<sub>2</sub>) salt bridge (known as 'ionic lock') between transmembrane helices 3 and 6; and (2) two active states, S<sub>3</sub> and S<sub>3'</sub>, as identified by binding of a G-protein-derived peptide. In contrast to a recent study of the β<sub>2</sub>-adrenergic receptor<sup>4</sup>, the present approach allowed identification of a second active state for A<sub>2A</sub>R. Addition of inverse agonist (ZM241385) increases the population of the inactive states, while full agonists (UK432097 or NECA) stabilize the active state, S<sub>3'</sub>, in a manner consistent with conformational selection. In contrast, partial agonist (LUF5834) and an allosteric modulator (HMA) exclusively increase the population of the S<sub>3</sub> state. Thus, partial agonism is achieved here by conformational selection of a distinct active state which we predict will have compromised coupling to the G protein. Direct observation of the conformational equilibria of ligand-dependent G-protein-coupled receptor and deduction of the underlying mechanisms of receptor activation will have wide-reaching implications for our understanding of the function of G-protein-coupled receptor in health and disease.

A myriad of signalling processes associated with vision, sensory response, neurotransmitter- and hormone-mediated response, inflammation, and cell homeostasis are governed by G-protein-coupled receptors (GPCRs), also called seven transmembrane helix (7TM) receptors. A<sub>2A</sub>R is a family A GPCR and an important drug target for treating inflammation, cancer, ischaemia reperfusion injury, sickle cell disease, diabetic nephropathy, infectious diseases, and neuronal disorders<sup>5</sup>. An understanding of the mechanism of GPCR activation and the representative conformational states is key to the drug design process. Our molecular perspective of activation is biased by X-ray crystallography, where the receptor is stabilized through thermostable mutants, fusion protein constructs, and appropriate ligands to obtain a single lowest-energy structure, often designated as either 'inactive' or 'active'. Using <sup>19</sup>F NMR and judiciously placed tags, we observed A<sub>2A</sub>R in a dynamic equilibrium between two inactive and two active states. The activation process can thus be viewed from the perspective of populations of key functional states, and the action of ligands on this conformational landscape through conformational selection.

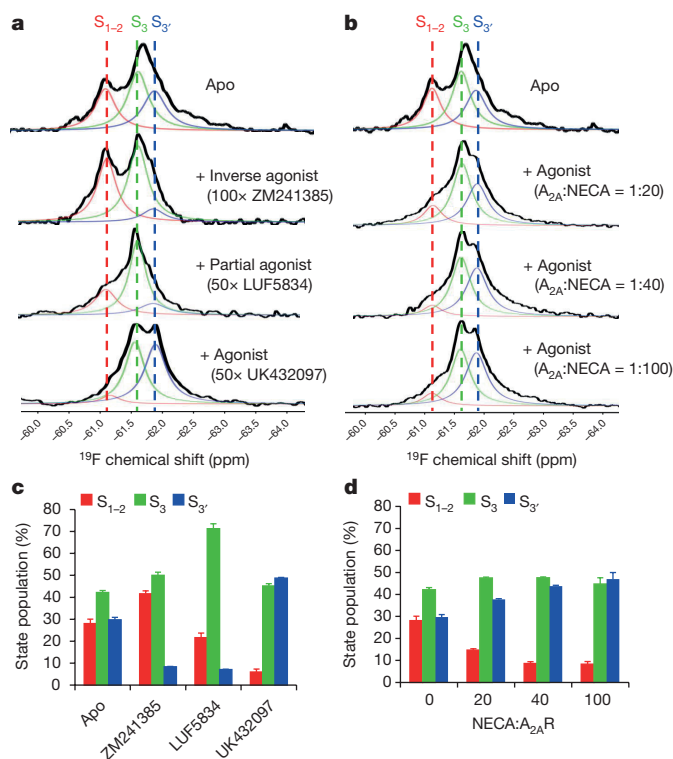
X-ray crystal structures of A<sub>2A</sub>R, stabilized either by inverse agonist or by agonist, suggest that receptor activation involves a rearrangement

of the 7TM bundle; that is, the inward shift of the intracellular part of TM7, a translation of TM3, and the formation of a bulge in TM5, in addition to an outward displacement and rotation of TM6 bringing together the intracellular ends of TM5 and TM6 (refs 6–8). Analogous observations were made for the β<sub>2</sub>-adrenergic receptor<sup>9</sup> (β<sub>2</sub>AR) and the light-activatable GPCR rhodopsin<sup>10–12</sup>, suggesting a common activation pathway (Extended Data Fig. 1). Via activation intermediates through which these TM domains rearrange, GPCRs form an increasingly larger crevice at the cytoplasmic side<sup>11</sup>, which is eventually large enough to harbour the key binding sites of interacting G protein and arrestin<sup>9,13</sup>. We used electron paramagnetic resonance (EPR) and NMR to identify labelling sites on TM5 and TM6. A <sup>19</sup>F NMR label at V229C on TM6 (Extended Data Figs 2 and 3) appeared to be ideal for monitoring activation of A<sub>2A</sub>R (the version used in this study is truncated after residue 317). In assessing conformational states and studying conformational exchange of GPCRs on the microsecond to millisecond timescale, both <sup>13</sup>C and <sup>19</sup>F NMR have proved useful<sup>4,14–18</sup>. In particular, <sup>19</sup>F NMR provides exquisite sensitivity to solvent exposure or side-chain packing, often revealing a wealth of conformations<sup>4,16,17</sup>.

A recent <sup>19</sup>F NMR study of β<sub>2</sub>AR identified four distinct states associated with receptor activation<sup>4</sup>. The apo form of β<sub>2</sub>AR was populated solely by two rapidly exchanging conformers corresponding to the 'ionic lock', a salt bridge between Arg131<sup>3,50</sup> on TM3 and Glu268<sup>6,30</sup> on TM6, either formed (S<sub>1</sub>) or broken (S<sub>2</sub>). An additional long-lived (lifetime τ = 660 ms) β<sub>2</sub>AR active state (S<sub>3</sub>), in slow exchange with S<sub>1</sub> and S<sub>2</sub>, was identified upon binding of agonist<sup>4</sup>. Further addition of a nanobody mimicking a G protein established another, fully active state (S<sub>4</sub>) of β<sub>2</sub>AR, deemed to be competent for signalling as concluded from the same maximally splayed cytoplasmic surface as in the β<sub>2</sub>AR•Gα<sub>s</sub> crystal structure<sup>4,9</sup>. Because neither of the two active states, S<sub>3</sub> and S<sub>4</sub>, could be detected in the ligand-free apo form of β<sub>2</sub>AR, it was not possible to distinguish between induced fit and conformational selection as models for β<sub>2</sub>AR activation.

In contrast to β<sub>2</sub>AR, the present <sup>19</sup>F NMR study revealed four states (two inactive and two active) associated with ligand-free apo A<sub>2A</sub>R<sup>6,8,19</sup> (Fig. 1 and Extended Data Figs 4 and 5). Owing to striking parallels with the previous study of β<sub>2</sub>AR, we have adopted a similar nomenclature for the states. The two inactive states S<sub>1</sub> and S<sub>2</sub> are in fast exchange on a millisecond timescale (Extended Data Fig. 4) and are represented by a single resonance, designated S<sub>1–2</sub>, which in analogy to β<sub>2</sub>AR flickers between an ionic lock stabilized (S<sub>1</sub>) and broken state (S<sub>2</sub>). Corresponding states are seen in A<sub>2A</sub>R crystal structures: a thermostabilized A<sub>2A</sub>R mutant with inverse agonist bound reveals an intact ionic lock between Arg102<sup>3,50</sup> and Glu228<sup>6,30</sup> (ref. 19), whereas A<sub>2A</sub>R structures with either antagonist<sup>6</sup> or agonist<sup>8</sup> bound show a broken ionic lock. Two upfield shifted resonances are associated with active states, S<sub>3</sub> and S<sub>3'</sub>, as identified by binding of G-protein-derived peptides (see below). In stark contrast to β<sub>2</sub>AR, the active states S<sub>3</sub> and S<sub>3'</sub> are already present in the A<sub>2A</sub>R apo form and their populations are

<sup>1</sup>Department of Chemistry, University of Toronto, UTM, 3359 Mississauga Road North, Mississauga, Ontario L5L 1C6, Canada. <sup>2</sup>Department of Biochemistry, University of Toronto, 1 King's College Circle, Toronto, Ontario M5S 1A8, Canada. <sup>3</sup>Department of Technical Biochemistry, University of Stuttgart, 31 Allmandring, Stuttgart, Baden-Württemberg, D-70569, Germany. <sup>4</sup>Department of Molecular Genetics, University of Toronto, 1 King's College Circle, Toronto, Ontario M5S 1A8, Canada.



**Figure 1 | Ligand-dependent A<sub>2A</sub>R state equilibria.** Three distinct resonances of <sup>19</sup>F-labelled A<sub>2A</sub>R-V229C are associated with inactive (S<sub>1-2</sub>, shown in red) and active states (S<sub>3</sub>, shown in green, and S<sub>3'</sub>, shown in blue), as a function of representative ligands. **a**, <sup>19</sup>F NMR spectra of the receptor in the apo form or in the presence of inverse agonist, partial agonist, or full agonist, respectively. **b**, <sup>19</sup>F NMR spectra of the receptor in the apo form and with increasing amounts of NECA agonist. **c**, Histogram obtained from spectral deconvolutions, comparing the relative populations of S<sub>1-2</sub>, S<sub>3</sub>, and S<sub>3'</sub> states. **d**, Histogram comparing the relative populations of states, S<sub>1-2</sub>, S<sub>3</sub>, and S<sub>3'</sub>, upon titration of the full agonist NECA to A<sub>2A</sub>R. Experiments were replicated at least three times from separately expressed and reconstituted samples. Details on the chemical shift referencing, line shape fitting procedure, and error analyses are provided in the Supplementary Information and Extended Data Figs 5 and 6.

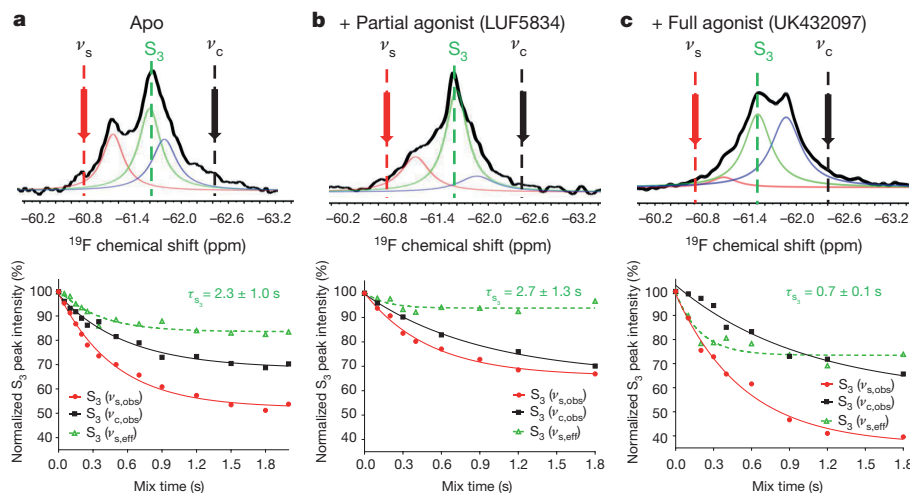
increased by the addition of partial agonist or full agonist, respectively. Addition of ligand merely alters the distribution of states in a manner consistent with conformational selection.

Figure 1 and Extended Data Fig. 6 show <sup>19</sup>F NMR spectra of 2-bromo-N-(4-(trifluoromethyl)phenyl) acetamide (BTFMA)-labelled A<sub>2A</sub>R-V229C as a function of representative ligands (inverse

agonist (ZM241385), partial agonist (LUF5834), and two full agonists (UK432097 and NECA)). Three resonances associated with states, S<sub>1-2</sub>, S<sub>3</sub>, and S<sub>3'</sub>, can be identified in all of the spectra. Addition of inverse agonist shifts the equilibrium towards the S<sub>1-2</sub> ensemble. Addition of the partial agonist LUF5834 stabilizes S<sub>3</sub>. The allosteric modulator HMA has the same effect on S<sub>3</sub> with the caveat that the resonance associated with S<sub>1-2</sub> appears to be exchange broadened (Extended Data Fig. 7). Finally, full agonists (UK432097 or NECA) shift the equilibrium towards S<sub>3'</sub>. The chemical shifts associated with S<sub>1-2</sub>, S<sub>3</sub>, and S<sub>3'</sub> are observed to be increasingly upfield, consistent with a corresponding increase in solvent exposure of the probe<sup>20</sup> and opening of the cytoplasmic crevice via rotation and translation of TM5 and TM6. The corresponding state populations are obtained directly from signal deconvolutions (Fig. 1a, b) and are provided as histograms (Fig. 1c, d).

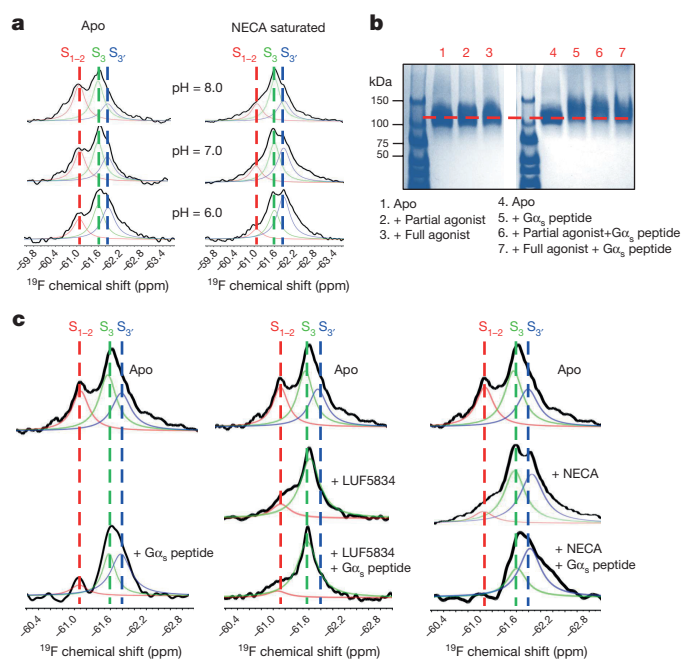
While inactive states S<sub>1</sub> and S<sub>2</sub> undergo exchange on a low millisecond timescale, exchange between the inactive state ensemble S<sub>1-2</sub> and the active states S<sub>3</sub> and S<sub>3'</sub> is of the order of 1 or 2 s, as shown by saturation transfer experiments (Fig. 2). In this case, the inactive ensemble can be selectively saturated by application of a low power pulse applied at a frequency,  $\nu_s$ , slightly downfield from the resonance associated with S<sub>1-2</sub>. By recording spectra as a function of the duration of the pulse, it is possible to determine the rate of exchange between S<sub>3</sub> and S<sub>1-2</sub>, or equivalently the lifetime of the S<sub>3</sub> intermediate state,  $\tau_{S_3}$ . Conversely, the lifetime of the inactive ensemble,  $\tau_{S_{1-2}}$ , can be determined by saturating the active states (S<sub>3</sub> and S<sub>3'</sub>) as described in Extended Data Fig. 8. Note that because of overlap between S<sub>3</sub> and S<sub>3'</sub>, it is difficult to measure their mutual exchange. The saturation transfer experiments (Fig. 2) reveal that the S<sub>3</sub> state is long-lived (1–3 s) for A<sub>2A</sub>R in the apo form or when bound to either inverse agonist or partial agonist. The addition of agonist (UK432097) appears to shorten the lifetime of the S<sub>3</sub> state, which may be a consequence of lowered barriers, and, hence, exchange between S<sub>3</sub> and both S<sub>1-2</sub> and the S<sub>3'</sub> states. The saturation transfer experiments are further consistent with a sequential transition S<sub>3'</sub> → S<sub>3</sub> → S<sub>1-2</sub> (Extended Data Fig. 8b, c).

A sequence of GPCR states where the receptor becomes gradually more active has been shown for the photoreceptor and GPCR rhodopsin<sup>3,11,12,21</sup>. According to this sequence of reaction steps, formation of the fully active receptor state is concomitant with a proton uptake from the aqueous environment to the conserved D(E)RY motif on TM3. We therefore recorded pH-dependent <sup>19</sup>F NMR spectra of BTFMA-labelled A<sub>2A</sub>R-V229C in the apo form and in the presence of saturating amounts of NECA agonist (Fig. 3). With decreasing pH, the population of states shifted towards the S<sub>3'</sub> state at the expense of S<sub>1-2</sub> and S<sub>3</sub>, as expected for a coupled equilibrium where the last transition from S<sub>3</sub> to S<sub>3'</sub> is pH-dependent. The pH-dependent population of the S<sub>3'</sub> state was more pronounced in the presence of NECA agonist (Fig. 3a). An analogy is seen with opsin (the apo form of rhodopsin), which is also



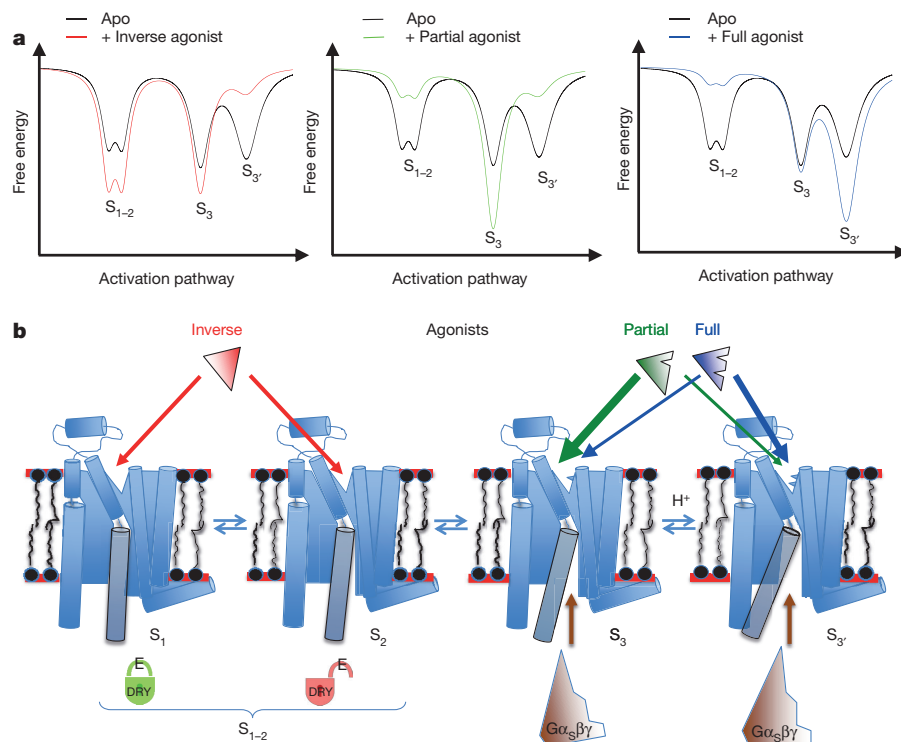
**Figure 2 | Ligand-induced effects on conformational state lifetimes.** <sup>19</sup>F NMR spectra of BTFMA-labelled A<sub>2A</sub>R-V229C and corresponding decay curves associated with S<sub>3</sub>, upon saturating S<sub>1-2</sub>. **a**, A<sub>2A</sub>R apo form. **b**, **c**, A<sub>2A</sub>R in the presence of saturating amounts of partial agonist (LUF5834; **b**) or full agonist (UK432097; **c**). To account for off-resonant saturation effects due to the pulse at a frequency,  $\nu_s$ , a control experiment was performed at a frequency,  $\nu_c$ , equidistant to the peak of interest. The effective decay curve (green dashed line) represents the approximate response of S<sub>3</sub> associated with selective saturation to S<sub>1-2</sub>.





**Figure 3 | Effect of pH and  $G\alpha_s$ -derived peptide on  $A_{2A}R$  conformational states.** **a**,  $^{19}F$  NMR spectra of BTFMA-labelled  $A_{2A}R$ -V229C at various pH values for the  $A_{2A}R$  apo form (left) and  $A_{2A}R$  saturated with NECA agonist (right). **b**, Native gel of BTFMA-labelled  $A_{2A}R$ -V229C in the apo form or in the presence of saturating amounts of partial agonist (LUF5834) or full agonist (NECA), respectively (lanes 1–3). The presence of a  $G\alpha_s$ -derived peptide causes a mobility shift (lanes 5–7). **c**,  $^{19}F$  NMR spectra of BTFMA-labelled  $A_{2A}R$ -V229C in either the apo form or in the presence of saturating amounts of partial agonist (LUF5834) or full agonist (NECA) in absence and presence of  $G\alpha_s$ -derived peptide. Ligand and peptide concentrations were  $50 \times$  LUF5834,  $100 \times$  NECA and  $50 \times$   $G\alpha_s$ -derived peptide, respectively, relative to the receptor concentration.

in a pH-dependent equilibrium between inactive and active states<sup>22</sup> and where stabilization of the active state is additionally facilitated by the presence of all-*trans*-retinal agonist<sup>12,21</sup>.



**Figure 4 | Model of the free energy landscape and corresponding model of  $A_{2A}$  receptor activation.**

**a**, The effects of inverse agonist, partial agonist, and full agonist on the state equilibria are illustrated in the free energy landscapes. The functional states ( $S_{1-2}$ ,  $S_3$ , and  $S_{3'}$ ) are characterized as sitting in deep free-energy wells, while undergoing relatively slow exchange. Ligands affect this landscape in a manner consistent with conformational selection. **b**, Binding of  $G\alpha_s\beta\gamma$  to apo  $A_{2A}R$  is enabled through the active state ensemble. Partial agonists and full agonists either stabilize  $S_3$  or  $S_{3'}$ , respectively. This gives rise to two levels of binding and activation of  $G\alpha_s\beta\gamma$ .

In Fig. 3b, c, we examine the effect of a peptide derived from the carboxyl (C)-terminal domain of the G-protein  $G\alpha_s$  (RVF NDARDIIQRMHLRQYELL)<sup>23</sup> on the equilibrium of  $A_{2A}R$  states.  $^{19}F$  NMR data and mobility shifts in native gels showed that the peptide is able to interact with the apo receptor and  $A_{2A}R$  saturated with partial agonist or full agonist. Addition of the peptide reduced the inactive state ensemble population and shifted the equilibrium towards the  $S_3$  and  $S_{3'}$  states, identifying both states as active as characterized by their capability to interact with the  $G\alpha_s$ -derived peptide.  $S_3$  and  $S_{3'}$  states have different conformations and thus may vary in their capacity to activate G protein. In the presence of saturating amounts of full agonist, addition of the  $G\alpha_s$  peptide resulted in a pronounced shift towards  $S_3$  and  $S_{3'}$ , whereas in the presence of partial agonist the  $S_3$  intermediate prevailed without population of  $S_{3'}$ . The spectra thus demonstrate that the  $G\alpha_s$  peptide is able to bind either  $S_3$  or  $S_{3'}$  states in a manner consistent with conformational selection. Moreover, a closer inspection of the apo spectrum suggests that the peptide preferentially binds to  $S_{3'}$  over  $S_3$ , which is not the case in the presence of a partial agonist. Rather, the partial agonist stabilizes the  $S_3$  state, and addition of  $G\alpha_s$ -derived peptide only reinforces this state. This probably directly relates to a reduced efficiency of binding and activation of the holo G protein when partial agonist stabilizes  $A_{2A}R$ .

The activation process associated with GPCRs is probably best understood in the case of visual rhodopsin with its covalently bound chromophore 11-*cis*-retinal<sup>11,12,21</sup>. Light absorption causes *cis/trans* isomerization and thus *in situ* conversion of an inverse agonist into an agonist. The fully active G-protein-interacting state develops sequentially through a series of metarhodopsin states which are in equilibrium and are stabilized by proton uptake. We find a remarkable similarity for  $A_{2A}R$  with inactive and active states, which find their counterparts in the rhodopsin activation scheme as proposed earlier<sup>21</sup>. The opsin apo form exists in a pH-dependent conformational equilibrium<sup>22</sup> and retinal uptake is suggested to occur via conformational selection<sup>24</sup>.

The current NMR data reaffirm the idea that key functional states simultaneously exist within a dynamic and 'loosely coupled' ensemble<sup>25</sup> of the unliganded receptor, as depicted in Fig. 4. Inactive and active states exchange slowly, as has been previously noted in studies of other GPCRs<sup>4,26,27</sup>. The corresponding high activation barriers probably



play a key role in regulation of signalling. Despite these barriers, basal activity of a receptor such as A<sub>2A</sub>R would be expected to occur owing to the presence of S<sub>3'</sub> and, presumably to a lesser extent, S<sub>3</sub>. An inverse agonist shifts the equilibrium towards the inactive ensemble, S<sub>1-2</sub>, and suppresses the basal population of active states.

The addition of partial agonist or full agonist further stabilizes the respective active states, consistent with the notion of conformational selection<sup>28</sup>, while it is also clear that ligands influence barrier heights associated with activation, as exemplified by the observation that HMA resulted in faster exchange between S<sub>1-2</sub> and S<sub>3</sub> (Extended Data Fig. 7). We note that 70% of the unliganded receptors adopt the active states, S<sub>3</sub> or S<sub>3'</sub>, in contrast to β<sub>2</sub>AR, where the receptor was biased towards the inactive ensemble and active states could only be detected with agonist and/or nanobody<sup>4</sup>. It may be that, in the absence of agonists, the active states associated with β<sub>2</sub>AR are either very weakly populated or are short-lived and therefore exchange-broadened to the point where they cannot yet be detected. The coexistence of S<sub>3</sub> and S<sub>3'</sub> in A<sub>2A</sub>R also underlies the concept of partial agonism, which refers to a pharmacological phenomenon where the addition of saturating amounts of a given agonist results in sub-maximal signalling or efficacy. There are two schools of thought as to how partial agonism might originate at a molecular level. Quite simply, a partial agonist may establish an unstable (short-lived) fully active state<sup>29</sup>. In this case, the peak associated with a partial agonist would be represented by a weighted average between resonances associated with the fully active and other active and inactive states and might also exhibit exchange broadening. The second possibility is that a partial agonist sub-optimally engages the orthosteric binding site such that the active conformation is simply not fully established (that is, S<sub>3</sub> in A<sub>2A</sub>R). The two distinct frequencies, which prevail for any ligand tested, imply the existence of two inherently stable states, whose relative populations are determined by the ligand and/or environment. The addition of partial agonist would result in a 'less open' A<sub>2A</sub>R conformation, leading to weaker allosteric coupling with the G protein, than that attained through the 'more open' S<sub>3'</sub> state. This notion that the conformation of the partial agonist stabilized state is not fully competent to engage with the G protein is reminiscent of a recent description of partial agonism in terms of the triggering of multiple switches in the receptor<sup>30</sup>.

We note that V229C was selected as a labelling site, as discussed in the Methods, to discriminate between active and inactive states, identified by crystallography, while minimizing expression losses, misfolding, or loss of function. A survey of other domains may reveal additional nuances associated with the activation process. We have been able to demonstrate that ligand binding to A<sub>2A</sub>R occurs through conformational selection rather than induced fit. This has important ramifications for drug design as it implies that any therapeutic compound would ideally favour a pre-existing receptor state. Accordingly, choice of the interacting state would then dictate pharmacology.

**Online Content** Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

**Received 21 November 2015; accepted 16 March 2016.**

**Published online 4 May 2016.**

- Weigl, T. R. & Paul, F. Conformational selection in protein binding and function. *Protein Sci.* **23**, 1508–1518 (2014).
- Nussinov, R., Ma, B. & Tsai, C. J. Multiple conformational selection and induced fit events take place in allosteric propagation. *Biophys. Chem.* **186**, 22–30 (2014).
- Deupi, X. & Kobilka, B. K. Energy landscapes as a tool to integrate GPCR structure, dynamics, and function. *Physiology* **25**, 293–303 (2010).
- Manglik, A. *et al.* Structural insights into the dynamic process of β<sub>2</sub>-adrenergic receptor signaling. *Cell* **161**, 1101–1111 (2015).
- Ruiz, M. L., Lim, Y. H. & Zheng, J. Adenosine A<sub>2A</sub> receptor as a drug discovery target. *J. Med. Chem.* **57**, 3623–3650 (2014).

- Jaakola, V. P. *et al.* The 2.6 angstrom crystal structure of a human A<sub>2A</sub> adenosine receptor bound to an antagonist. *Science* **322**, 1211–1217 (2008).
- Lebon, G. *et al.* Agonist-bound adenosine A<sub>2A</sub> receptor structures reveal common features of GPCR activation. *Nature* **474**, 521–525 (2011).
- Xu, F. *et al.* Structure of an agonist-bound human A<sub>2A</sub> adenosine receptor. *Science* **332**, 322–327 (2011).
- Rasmussen, S. G. *et al.* Crystal structure of the β<sub>2</sub> adrenergic receptor–Gs protein complex. *Nature* **477**, 549–555 (2011).
- Choe, H. W. *et al.* Crystal structure of metarhodopsin II. *Nature* **471**, 651–655 (2011).
- Hofmann, K. P. *et al.* A G protein-coupled receptor at work: the rhodopsin model. *Trends Biochem. Sci.* **34**, 540–552 (2009).
- Ernst, O. P. *et al.* Microbial and animal rhodopsins: structures, functions, and molecular mechanisms. *Chem. Rev.* **114**, 126–163 (2014).
- Kang, Y. *et al.* Crystal structure of rhodopsin bound to arrestin by femtosecond X-ray laser. *Nature* **523**, 561–567 (2015).
- Kofuku, Y. *et al.* Efficacy of the β<sub>2</sub>-adrenergic receptor is determined by conformational equilibrium in the transmembrane region. *Nature Commun.* **3**, 1045 (2012).
- Nygaard, R. *et al.* The dynamic process of β<sub>2</sub>-adrenergic receptor activation. *Cell* **152**, 532–542 (2013).
- Liu, J. J., Horst, R., Katritch, V., Stevens, R. C. & Wüthrich, K. Biased signaling pathways in β<sub>2</sub>-adrenergic receptor characterized by <sup>19</sup>F-NMR. *Science* **335**, 1106–1110 (2012).
- Kim, T. H. *et al.* The role of ligands on the equilibria between functional states of a G protein-coupled receptor. *J. Am. Chem. Soc.* **135**, 9465–9474 (2013).
- Sounier, R. *et al.* Propagation of conformational changes during μ-opioid receptor activation. *Nature* **524**, 375–378 (2015).
- Doré, A. S. *et al.* Structure of the adenosine A<sub>2A</sub> receptor in complex with ZM241385 and the xanthines XAC and caffeine. *Structure* **19**, 1283–1293 (2011).
- Sykes, B. D., Weingarten, H. I. & Schlesinger, M. J. Fluorotyrosine alkaline phosphatase from *Escherichia coli*: preparation, properties, and fluorine-19 nuclear magnetic resonance spectrum. *Proc. Natl Acad. Sci. USA* **71**, 469–473 (1974).
- Okada, T., Ernst, O. P., Palczewski, K. & Hofmann, K. P. Activation of rhodopsin: new insights from structural and biochemical studies. *Trends Biochem. Sci.* **26**, 318–324 (2001).
- Vogel, R. & Siebert, F. Conformations of the active and inactive states of opsin. *J. Biol. Chem.* **276**, 38487–38493 (2001).
- Mazzoni, M. R. *et al.* A Gα<sub>s</sub> carboxyl-terminal peptide prevents G<sub>s</sub> activation by the A<sub>2A</sub> adenosine receptor. *Mol. Pharmacol.* **58**, 226–236 (2000).
- Schafer, C. T. & Farrens, D. L. Conformational selection and equilibrium governs the ability of retinals to bind opsin. *J. Biol. Chem.* **290**, 4304–4318 (2015).
- Dror, R. O. *et al.* Pathway and mechanism of drug binding to G-protein-coupled receptors. *Proc. Natl Acad. Sci. USA* **108**, 13118–13123 (2011).
- Bockenhauer, S., Fürstenberg, A., Yao, X. J., Kobilka, B. K. & Moerner, W. E. Conformational dynamics of single G protein-coupled receptors in solution. *J. Phys. Chem. B* **115**, 13328–13338 (2011).
- Vafabakhsh, R., Levitz, J. & Isacoff, E. Y. Conformational dynamics of a class C G-protein-coupled receptor. *Nature* **524**, 497–501 (2015).
- Kumar, S., Ma, B., Tsai, C. J., Sinha, N. & Nussinov, R. Folding and binding cascades: dynamic landscapes and population shifts. *Protein Sci.* **9**, 10–19 (2000).
- Lape, R., Colquhoun, D. & Sivilotti, L. G. On the nature of partial agonism in the nicotinic receptor superfamily. *Nature* **454**, 722–727 (2008).
- Ahuja, S. & Smith, S. O. Multiple switches in G protein-coupled receptor activation. *Trends Pharmacol. Sci.* **30**, 494–502 (2009).

**Supplementary Information** is available in the online version of the paper.

**Acknowledgements** This work was supported by the Natural Sciences and Engineering Research Council of Canada, research discovery award grant number 261980 (to R.S.P.) and the Canada Excellence Research Chair Program (to O.P.E., who is the Anne and Max Tanenbaum Chair in Neuroscience at the University of Toronto). We thank T. Kobayashi and R. Grishammer for providing plasmids with A<sub>2A</sub>R sequence. We thank J. Wells, S. Larda, and F. Huang from the University of Toronto, as well as S. Furness, B. K. Kobilka, and R. Sunahara for their suggestions and comments.

**Author Contributions** L.Y., O.P.E., and R.S.P. designed the research. L.Y. performed the molecular biology work, generated high-yield transformants, and optimized receptor expression and purification. L.Y. also performed NMR and EPR labelling, NMR experiments, and analysed spectroscopy data. N.V.E. performed and analysed data from EPR experiments. M.Z. assisted with cell culture and receptor purification. R.S.P., L.Y., and O.P.E. prepared the manuscript. O.P.E. and R.S.P. supervised the project.

**Author Information** Reprints and permissions information is available at [www.nature.com/reprints](http://www.nature.com/reprints). The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to O.P.E. ([oliver.ernst@utoronto.ca](mailto:oliver.ernst@utoronto.ca)) or R.S.P. ([scott.prosser@utoronto.ca](mailto:scott.prosser@utoronto.ca)).

## METHODS

No statistical methods were used to predetermine sample size. The experiments were not randomized. The investigators were not blinded to allocation during experiments and outcome assessment.

**Plasmid construction and transformation.** The full-length human  $A_{2A}R$  gene, originating from construct pPIC9K\_ADORA2A<sup>31</sup>, was provided by T. Kobayashi. The construct M-TEV-hA2ARTr316-H10 (with a TEV protease cleavage site insert) was engineered from the construct M-hA2ARTr316-H10, which was provided by R. Grishammer<sup>32</sup>. A gene fragment,  $F_{\alpha}$ -Factor-Flag-TEV-A2ARTr316-H10, with components derived either from pPIC9K\_ADORA2A or from M-TEV-hA2ARTr316-H10 was amplified by fusion PCR with primers listed in Extended Data Table 1. pPIC9K\_ADORA2A and  $F_{\alpha}$ -Factor-Flag-TEV-A2ARTr316-H10 were digested with BamHI-HF and NotI-HF (New England Biolabs) restriction enzymes. The isolated  $F_{\alpha}$ -Factor-Flag-TEV-A2ARTr316-H10 fragment was subcloned into the pPIC9K plasmid to generate the new plasmid pPIC9K\_ $F_{\alpha}$ -Factor-Flag-TEV-A2ARTr316-H10. The construct pPIC9K\_ $F_{\alpha}$ -Factor-Flag-TEV-A2ARTr316-H10\_V229C containing the V229C mutation was generated using a QuikChange Lightning Site-Directed Mutagenesis Kit (Agilent Technologies) with primers listed in Extended Data Table 1. All constructs were sequenced by a local DNA sequencing facility (The Centre for Applied Genomics, Sick Kids Hospital, Toronto, Canada) with the AOX1 primer pair of PF<sub>AOX1</sub> and PR<sub>AOX1</sub>, listed in Extended Data Table 1. The proteins resulting from plasmids pPIC9K\_ $F_{\alpha}$ -Factor-Flag-TEV-A2ARTr316-H10 and pPIC9K\_ $F_{\alpha}$ -Factor-Flag-TEV-A2ARTr316-H10\_V229C were designated as  $A_{2A}R$  (Pro2 to Ala317) and  $A_{2A}R$ -V229C (Pro2 to Ala317 including the V229C mutation), respectively, on the basis of their correspondence to the wild-type sequence. Freshly prepared competent cells of a strain of *Pichia pastoris* SMD 1163 ( $\Delta his4 \Delta pep4 \Delta prb1$ , Invitrogen) were electro-transformed with PmeI-HF (New England Biolabs) linearized plasmids using a Gene Pulser II (Bio-Rad). High-copy clone selection was performed as previously described<sup>33</sup>, and a high-yield construct was then screened by an immunoblotting assay for further expression.

**Receptor expression, purification and labelling.** A pre-cultured single colony on YPD (1% (w/v) yeast extract, 2% (w/v) peptone and 2% (w/v) glucose) plates containing 0.1 mg ml<sup>-1</sup> G418 was inoculated into 4 ml YPD medium and cultured at 30 °C for 12 h, then transferred into 200 ml BMGY medium (1% (w/v) yeast extract, 2% (w/v) peptone, 1.34% (w/v) YNB (yeast nitrogen base) without amino acids, 0.00004% (w/v) biotin, 1% (w/v) glycerol, 0.1 M PB (phosphate buffer) at pH 6.5) and cultured at 30 °C for another 24 h with an absorbance  $A_{595nm}$  in the range 2–6. The cell pellets were spun down at 4,000g for 5 min and were then resuspended in 1 l of BMMY medium (1% (w/v) yeast extract, 2% (w/v) peptone, 1.34% (w/v) YNB without amino acids, 0.00004% (w/v) biotin, 0.5% (w/v) methanol, 0.1 M phosphate buffer at pH 6.5, 0.04% (w/v) histidine and 3% (v/v) DMSO, 10 mM theophylline) at 20 °C. Methanol (0.5% (v/v)) was added every 18 h. Sixty hours after induction by methanol, cells were harvested for purification.

The cell pellets were collected by centrifugation at 4,000g for 10 min, and washed twice with washing buffer (50 mM HEPES, 10% glycerol, pH 7.4) before addition of breaking buffer (50 mM HEPES, pH 7.4, 100 mM NaCl, 2 mM EDTA, 10% glycerol, 100 U Zymolyase, 100  $\mu$ M theophylline). The sample was kept at room temperature (20 °C) for 1 h before disruption by vortexing for 2 h at 4 °C. Intact cells and cell debris were separated from the membrane suspension by low speed centrifugation (8,000g) for 30 min. The supernatant was collected and centrifuged at 100,000g for 1 h, and the precipitated cell membrane was then immediately dissolved in membrane lysis buffer (50 mM HEPES, pH 7.4, 100 mM NaCl, 1% MNG-3 (lauryl maltose neopentyl glycol) and 0.2% CHS (cholesteryl hemisuccinate), 100  $\mu$ M theophylline, and 20 mM imidazole) under continuous agitation for 1–2 h at 4 °C until the membrane was dissolved. Subsequently, Talon resin (Clontech) was added to the solubilized membranes and incubated for at least 2 h or overnight under gentle agitation.

The  $A_{2A}R$ -bound Talon resin was washed twice with 50 mM HEPES buffer, pH 7.4, containing 100 mM NaCl, 0.1% MNG-3, and 0.02% CHS and resuspended in the same buffer, followed by addition of 100  $\mu$ M TCEP reducing agent and incubation for 20 min. TCEP was washed out immediately by two rinsing steps with a buffer made of 50 mM HEPES, pH 7.4, 100 mM NaCl, 0.1% MNG-3, and 0.02% CHS. The  $A_{2A}R$ -bound Talon resin was then resuspended in buffer made of 50 mM HEPES, pH 7.4, 100 mM NaCl, 0.1% MNG-3, and 0.02% CHS, and combined with 10- to 20-fold excess of the NMR label (2-bromo-N-(4-(trifluoromethyl)phenyl)acetamide, BTFMA)<sup>4,34</sup> or EPR label (3-(2-iodoacetamido)-PROXYL) in the presence of nitrogen and under gentle agitation overnight at 4 °C. At the same time, 20  $\mu$ l of TEV enzyme was added to remove the  $A_{2A}R$  amino (N)-terminal tag. Another aliquot of NMR label was then added and incubated for an additional 6 h to ensure complete labelling. After the labelling and removal of the N-terminal tag was complete,

the  $A_{2A}R$ -bound Talon resin was extensively washed in a disposable column with buffer containing 50 mM HEPES, pH 7.4, 100 mM NaCl, 0.1% MNG-3, and 0.02% CHS, and apo  $A_{2A}R$  was then eluted from the Talon resin with 50 mM HEPES, pH 7.4, 100 mM NaCl, 0.1% MNG-3, and 0.02% CHS, 250 mM imidazole and concentrated to a volume of 5 ml. NaCl and imidazole in the sample were then removed by dialysis against 100 ml of 50 mM HEPES, pH 7.4, 0.1% MNG-3, 0.02% CHS for 3 h. The XAC-agarose gel (antagonist xanthine amine congener (XAC) conjugated to Affi-Gel 10 resin) and  $A_{2A}R$  were then incubated together for 2 h under gentle agitation. Functional  $A_{2A}R$  was eluted with 50 mM HEPES, pH 7.4, 0.1% MNG-3, 0.02% CHS, 100 mM NaCl, 20 mM theophylline. The eluted samples were concentrated to 20 ml by centrifugal filtration (molecular weight cut-off 3.5 kDa), and Talon resin was added and incubated under gentle agitation for another 2 h to bind functional  $A_{2A}R$ . Functional  $A_{2A}R$  bound to the resin was washed extensively with buffer containing 50 mM HEPES, pH 7.4, 100 mM NaCl, 0.1% MNG-3, 0.02% CHS, and 20 mM imidazole, to remove all theophylline. Then, functional apo  $A_{2A}R$  was eluted with elution buffer (50 mM HEPES, pH 7.4, 100 mM NaCl, 0.1% MNG-3 and 0.02% CHS, 250 mM imidazole) and the sample was dialysed to remove imidazole and concentrated for NMR or EPR.

**Choosing labelling sites based on X-ray crystal structures.** PROSHIFT software<sup>35</sup> ([http://www.meilerlab.org/index.php/servers/show?s\\_id=9](http://www.meilerlab.org/index.php/servers/show?s_id=9)) was used to predict C $\alpha$  chemical shift differences between the crystal structures of the active NECA-bound state (Protein Data Bank (PDB) accession number 2YDV) and the inactive ZM241385-bound state (PDB accession number 3EML). For the present study, we focused on differences associated with transmembrane domains TM3, TM5, and TM6, in an effort to characterize different active states.

**Double electron–electron resonance and continuous wave EPR experiments.** Site V229C<sup>6,31</sup> was spin-labelled with 3-(2-iodoacetamido)-PROXYL to generate a paramagnetic nitroxide side chain. X-band continuous wave (CW)-EPR data of the spin-labelled apo  $A_{2A}R$  were acquired using a Bruker ELEXSYS E500 CW-EPR spectrometer coupled to an ER 4123D dielectric resonator. The field sweep for data collection was 100-G and modulation amplitude was 2-G. Data sets were typically averages of 30–50 scans. Double electron–electron resonance (DEER) spectroscopy was used to verify that only a single EPR active spin label was attached to  $A_{2A}R$ . The identical sample as in the continuous wave experiment was used to collect Q-band DEER data using a Bruker ELEXSYS E580 spectrometer. Data were analysed using the program 'LongDistances' developed by C. Altenbach, which is available for download at <http://www.biochemistry.ucla.edu/biochem/Faculty/Hubbell>.

**NMR experiments.** NMR samples typically consisted of 250  $\mu$ l volumes of 50–200  $\mu$ M <sup>19</sup>F-labelled  $A_{2A}R$ -V229C in 50 mM HEPES buffer and 100 mM NaCl, doped with 10% D<sub>2</sub>O. The receptor was stabilized in 0.1% MNG-3 and 0.02% CHS. All <sup>19</sup>F NMR experiments were performed on a 600 MHz Varian Inova spectrometer using a cryogenic triple resonance probe, with the high-frequency channel tuneable either to <sup>1</sup>H or to <sup>19</sup>F. Typical experimental setup included a 23  $\mu$ s 90° excitation pulse, an acquisition time of 200 ms, a spectral width of 15 kHz, and a repetition time of 1 s. Most spectra were acquired with 15,000 scans, which provided a signal-to-noise ratio of roughly 100. Processing typically involved zero filling, and exponential apodization equivalent to 15 Hz line broadening.

**T<sub>2</sub> measurements.** <sup>19</sup>F-labelled apo  $A_{2A}R$ -V229C (200  $\mu$ M; comprising amino acids 2–317) in the buffer as described above was used for measurements of transverse relaxation time ( $T_2$ ) by a CPMG  $T_2$  pulse sequence, using a refocusing period of 133  $\mu$ s, with a total transverse magnetization evolution time of 0.4, 0.8, 1.2, 1.6, 2.0, 2.4, 2.8, 3.2, and 3.6 ms.

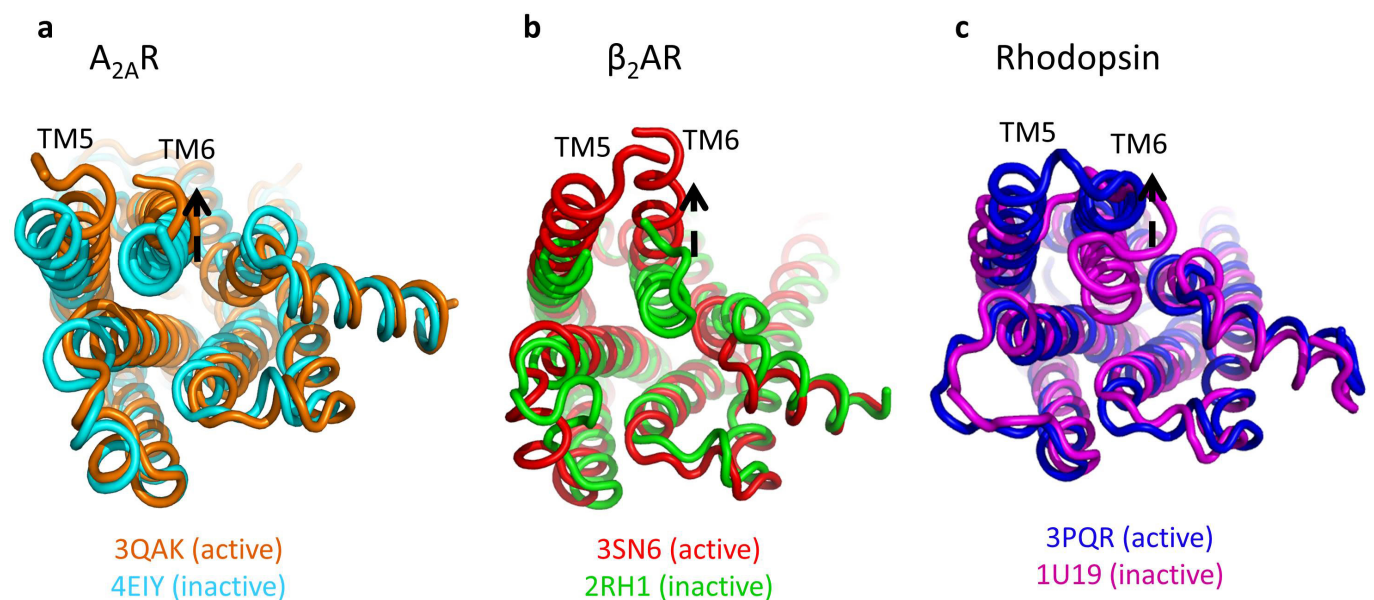
**<sup>19</sup>F saturation transfer experiments.** To investigate slow exchange between resolved states, <sup>19</sup>F chemical exchange saturation transfer NMR experiments were performed, in which a series of continuous-wave irradiation pulses were applied both at an on-resonance frequency ( $\nu_A$ ) and at an off-resonance frequency ( $\nu_C$ ), to assess chemical exchange during steady-state saturation and off-resonant saturation effects, as shown in Fig. 2. Upon saturating the resonance associated with state B, the ideal magnetization response of A may be described by the formula<sup>36</sup>

$$M_t^A = M_0^A \left( \frac{k_{AB}}{\rho_A + k_{AB}} \exp[-\tau(\rho_A + k_{AB})] + \frac{\rho_A}{\rho_A + k_{AB}} \right),$$

assuming off-resonant effects are accounted for. Note that both the exchange rate constants,  $k_{AB}$ , and the longitudinal relaxation rate of spin A,  $\rho_A$ , can in principle be calculated from a fit of the above equation to the experimental data. Accordingly, the lifetime  $\tau_A$  can be calculated from  $\tau = 1/k_{AB}$ . All fits were performed using Gnuplot (<http://www.gnuplot.info>).

- André, N. et al. Enhancing functional production of G protein-coupled receptors in *Pichia pastoris* to levels required for structural studies via a single expression screen. *Protein Sci.* **15**, 1115–1126 (2006).

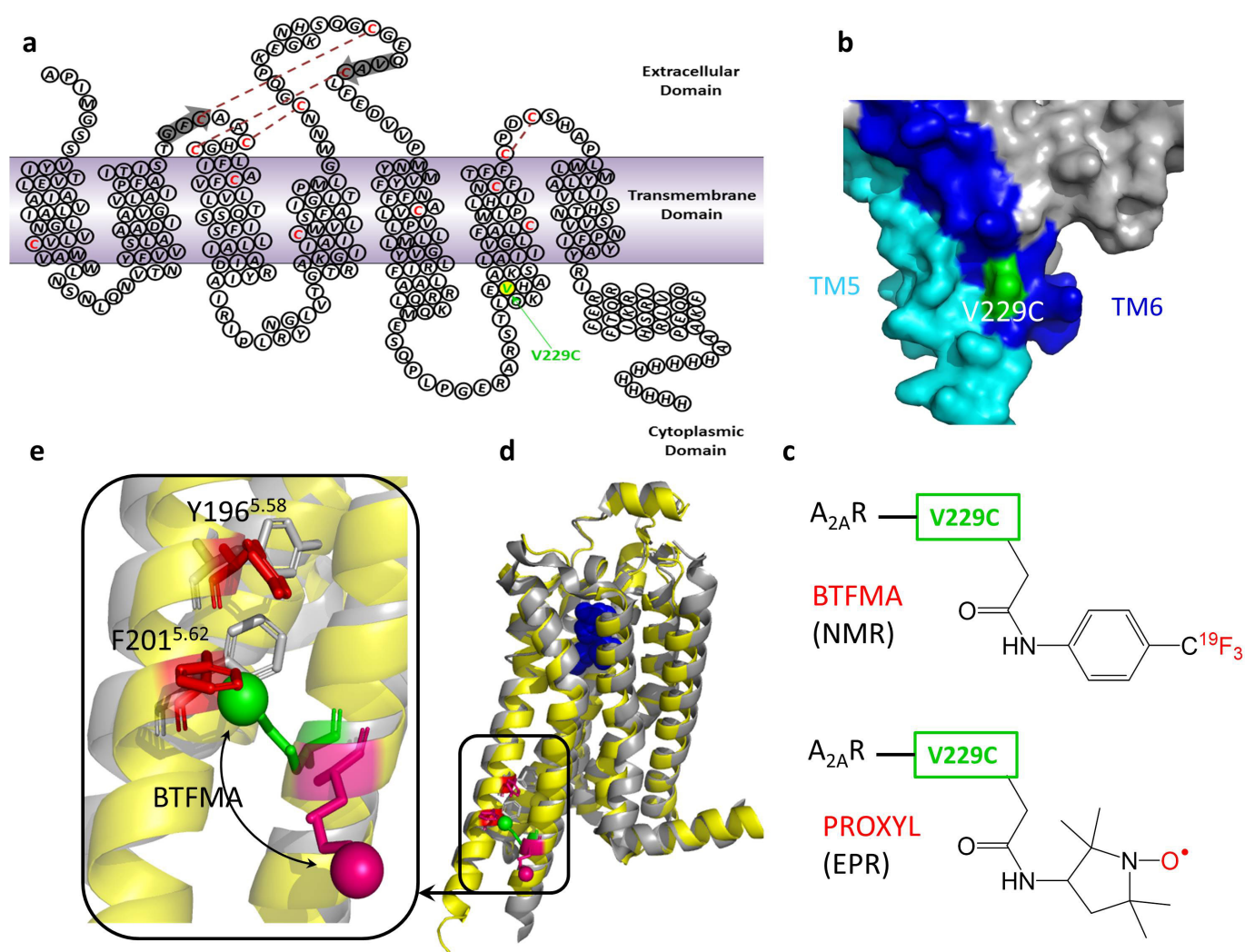
32. Weiß, H. M. & Grisshammer, R. Purification and characterization of the human adenosine A<sub>2a</sub> receptor functionally expressed in *Escherichia coli*. *Eur. J. Biochem.* **269**, 82–92 (2002).
33. Scorer, C. A., Clare, J. J., McCombie, W. R., Romanos, M. A. & Sreekrishna, K. Rapid selection using G418 of high copy number transformants of *Pichia pastoris* for high-level foreign gene expression. *Bio/Technology* **12**, 181–184 (1994).
34. Ye, L., Larda, S. T., Frank Li, Y. F., Manglik, A. & Prosser, R. S. A comparison of chemical shift sensitivity of trifluoromethyl tags: optimizing resolution in <sup>19</sup>F NMR studies of proteins. *J. Biomol. NMR* **62**, 97–103 (2015).
35. Meiler, J. PROSHIFT: protein chemical shift prediction using artificial neural networks. *J. Biomol. NMR* **26**, 25–37 (2003).
36. Helgstrand, M., Härd, T. & Allard, P. Simulations of NMR pulse sequences during equilibrium and non-equilibrium chemical exchange. *J. Biomol. NMR* **18**, 49–63 (2000).
37. Ballesteros, J. A. & Weinstein, H. Integrated methods for the construction of three-dimensional models and computational probing of structure-function relations in G protein-coupled receptors. *Methods Neurosci.* **25**, 366–428 (1995).



**Extended Data Figure 1 | Comparison of inactive and active GPCR crystal structures.** **a**, Inactive state  $A_{2A}R$  (cyan, inverse agonist ZM241385 bound, PDB accession number 4EIY) and active state  $A_{2A}R$  (brown, agonist UK432097 bound, PDB accession number 3QAK). **b**, Inactive state  $\beta_2AR$  (green, inverse agonist carazolol bound, PDB accession number 2RH1) and active state  $\beta_2AR$  (red, agonist (8-[(1R)-2-[[1,1-dimethyl-2-

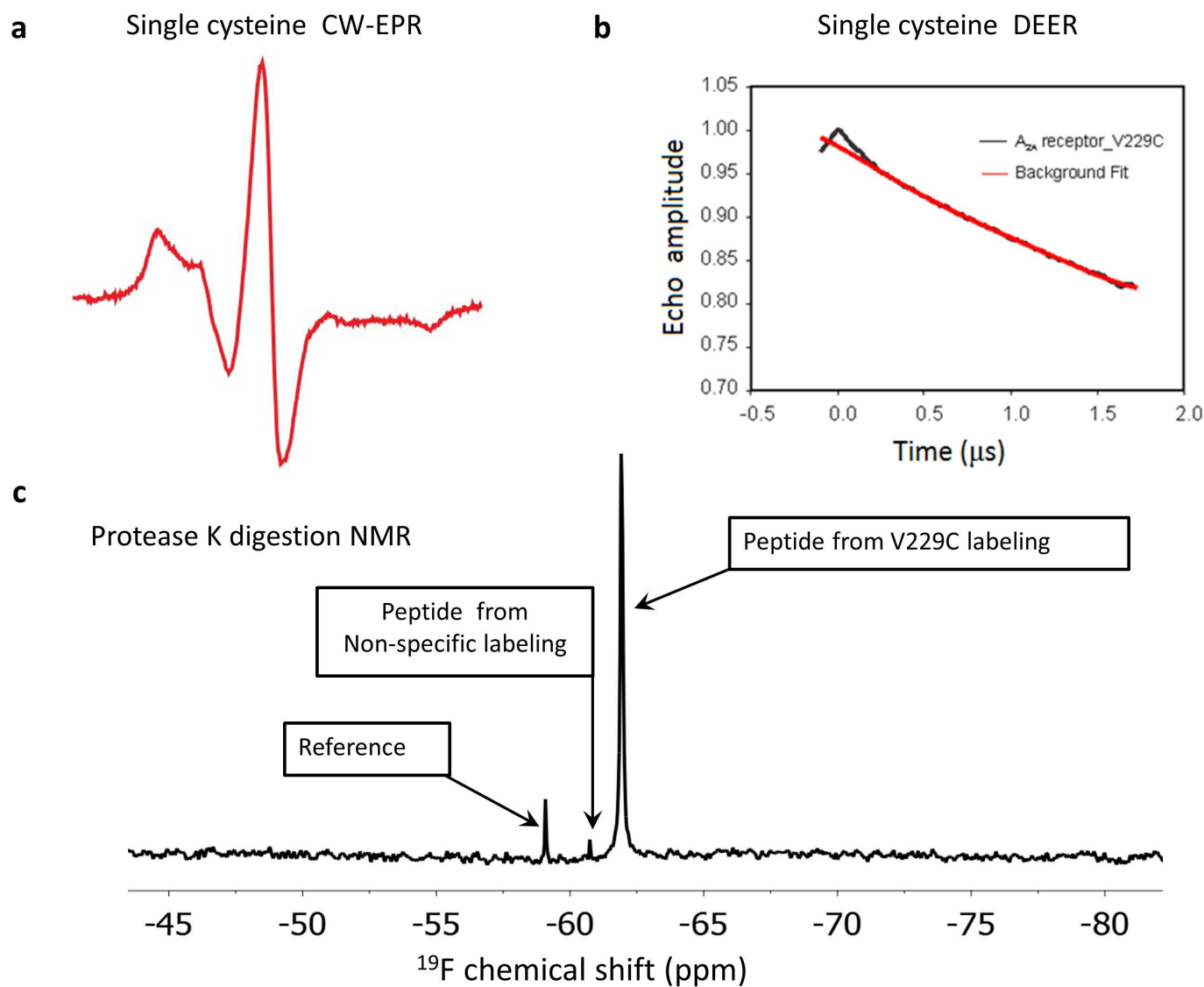
(2-methylphenyl)ethyl] amino)-1-hydroxyethyl]-5-hydroxy-2H-1,4-benzoxazin-3(4H)-one) bound, PDB accession number 3SN6). **c**, Inactive rhodopsin (purple, inverse agonist 11-*cis*-retinal bound, PDB accession number 1U19) and active metarhodopsin II (blue, agonist all-*trans*-retinal bound, PDB accession number 3PQR).





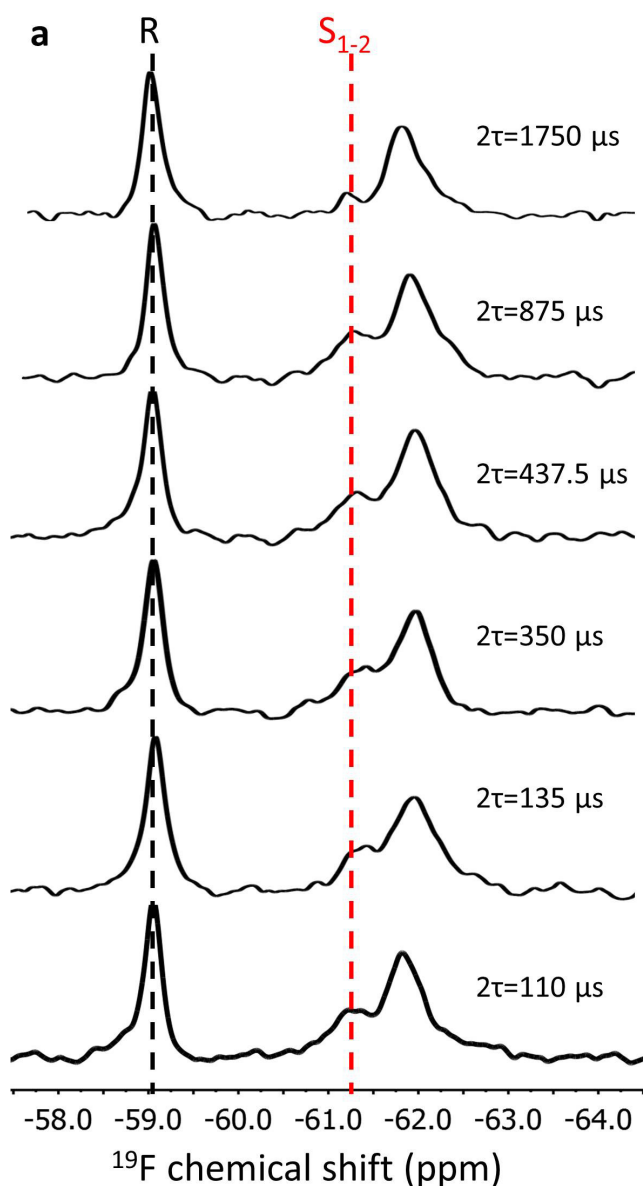
**Extended Data Figure 2 | Secondary structure and topology of C-terminally truncated A<sub>2A</sub>R-V229C.** Residues 2–317 of A<sub>2A</sub>R are preceded by an Ala residue resulting from TEV protease cleavage, and are succeeded by an Ala–His<sub>10</sub> sequence. A<sub>2A</sub>R was expressed in *P. pastoris* (SMD1163 strain) through genomic integration of a pPIC9K vector with a leader sequence consisting of  $\alpha$ -Factor, Flag tag (DYKDDDDK), and a TEV protease recognition domain (SNNNNNNNNNGLGENLYFQGA). During the secretion process, the signal peptide of the  $\alpha$ -Factor gets cleaved and the domain associated with the Flag tag and TEV recognition domain is removed by TEV protease. **a**, The truncated wild-type receptor used in this study contains all four native disulfide bonds and six buried cysteine residues (indicated in red), none of which were perturbed by the labelling process, which was specific for the introduced (solvent-exposed)

cysteine residue V229C<sup>6,31</sup> (shown in green with yellow background; the superscript refers to the Ballesteros–Weinstein numbering<sup>37</sup>). **b**, A surface map suggests V229C (green, solvent exposed) should be fully labelled without perturbing the receptor. **c**, Structures of protein-attached labels for NMR (BTFMA; 2-bromo-*N*-(4-(trifluoromethyl)phenyl)acetamide) and EPR (PROXYL; 3-(2-iodoacetamido)-PROXYL) analysis. **d, e**, Location and topology of the labelling site associated with V229C for both the inverse agonist (inactive, grey) and agonist-bound (active, yellow) states (PDB accession numbers 4EIY and 3QAK). Two rotamers of the BTFMA label are indicated in green and purple (the phenyl moiety is shown as a sphere). Note that the size of the tags is slightly larger than that depicted in the figure. The environment around the tag is predicted to differ for inactive and active states of the receptor.

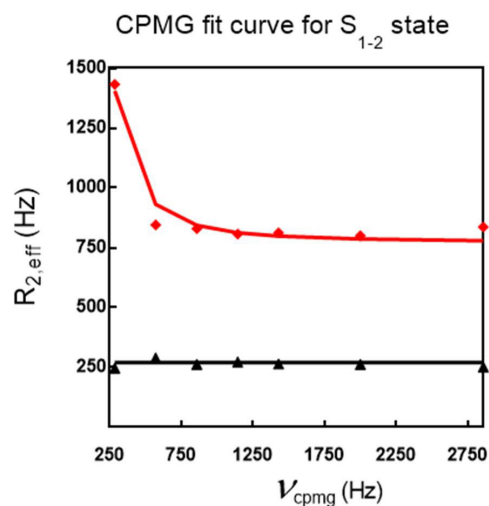


**Extended Data Figure 3 | Labelling efficiency of A<sub>2A</sub>R-V229C.** a, Single cysteine CW-EPR spectrum of 50  $\mu$ M apo A<sub>2A</sub>R-V229C receptor, labelled with a PROXYL spin-label and reconstituted into MNG-3 detergent

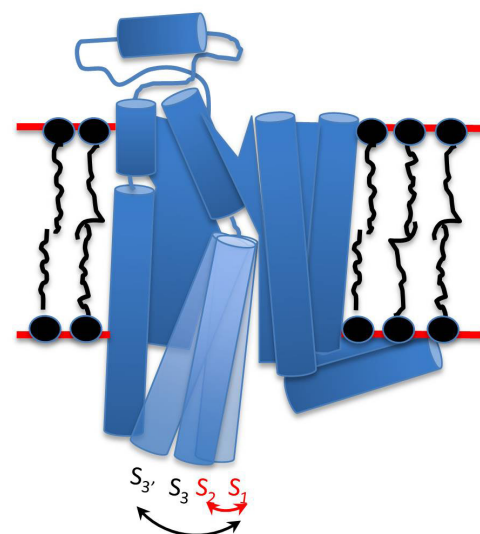
micelles. b, DEER measurement of 50  $\mu$ M PROXYL spin-labelled apo A<sub>2A</sub>R-V229C receptor. c,  $^{19}\text{F}$  NMR spectra of protease-K-digested  $^{19}\text{F}$ -labelled A<sub>2A</sub>R-V229C, showing one dominant peak.



b

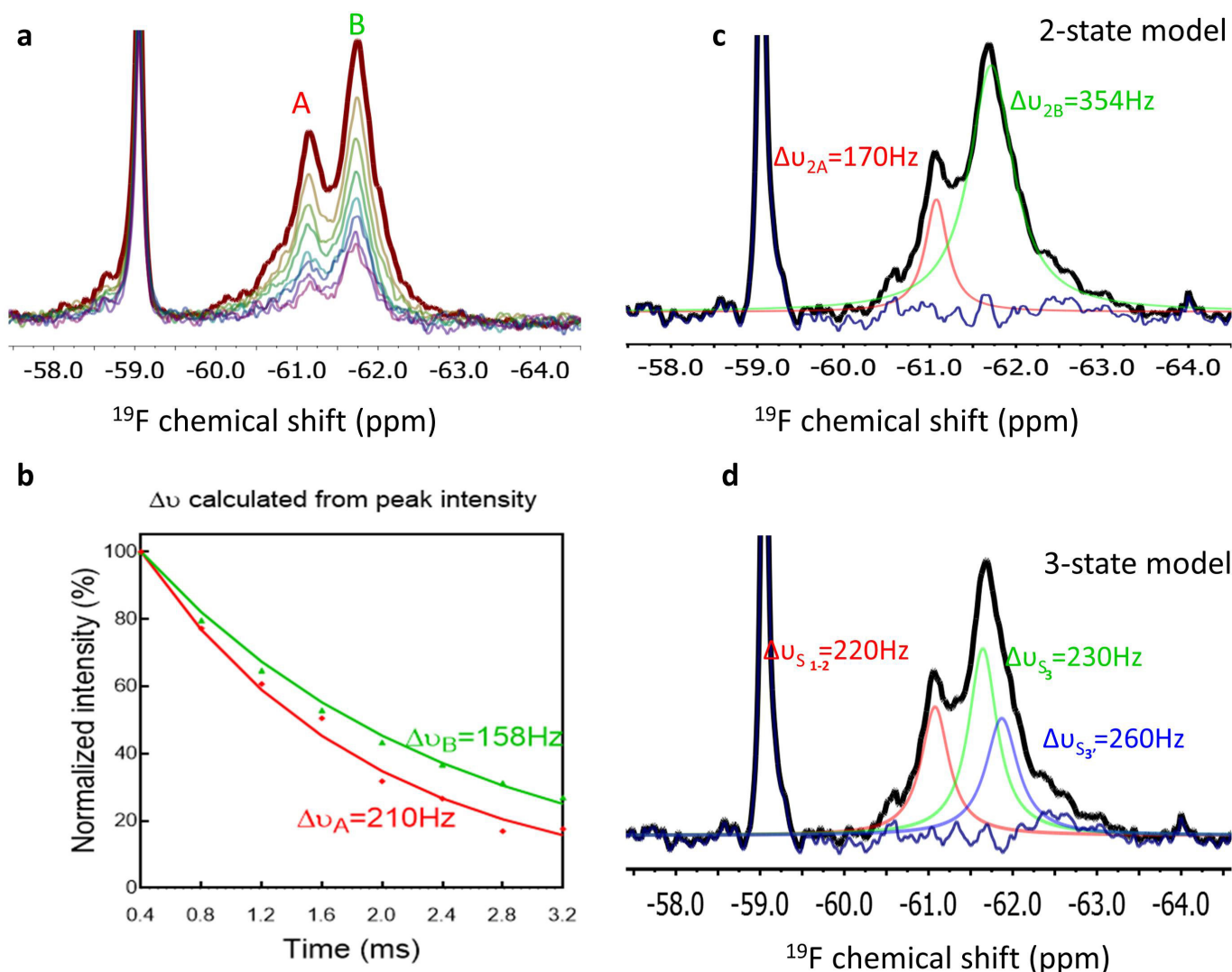


c



**Extended Data Figure 4 | Car-Purcell-Meiboom-Gill (CPMG) relaxation dispersion experiment to evaluate dynamics of  $S_{1-2}$ .** **a**,  $^{19}\text{F}$  NMR CPMG relaxation series of  $^{19}\text{F}$ -labelled apo  $A_{2A}\text{R-V229C}$ . Each spectrum was acquired using 10,000 scans with a constant  $T_2$ -refocusing period of 3.5 ms. The spectra in the relaxation series were recorded with different refocusing frequencies (that is, different periods between the refocusing pulses as indicated above, representative of three experiments).

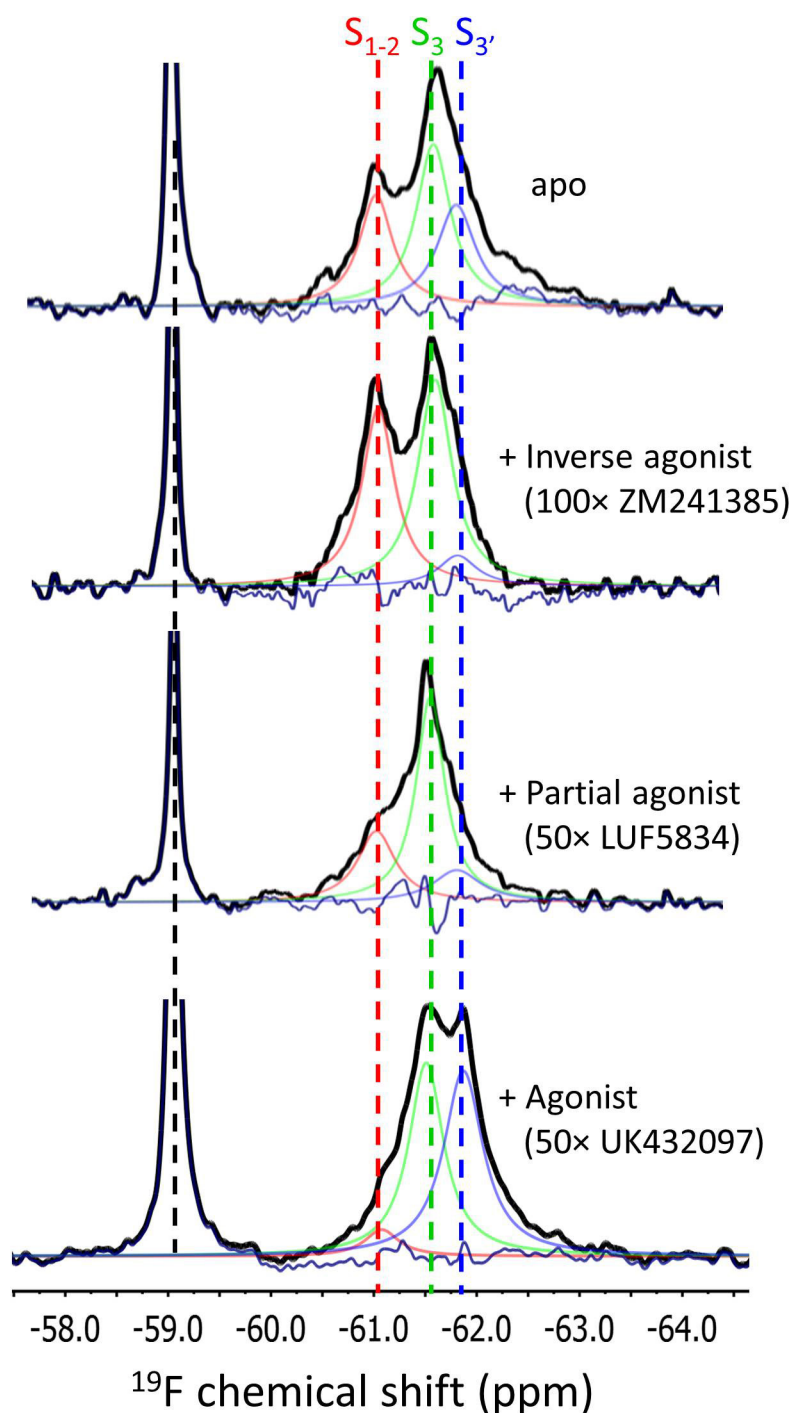
The sample consisted of  $200\ \mu\text{M}$   $^{19}\text{F}$ -labelled apo  $A_{2A}\text{R-V229C}$  in 50 mM HEPES buffer (pH 7.4) and 100 mM NaCl. **b**, CPMG curve for the  $S_{1-2}$  peak (red diamonds) and reference peak (black triangles).  $S_{1-2}$  undergoes millisecond timescale exchange while the reference peak exhibits no dispersion. **c**, Cartoon illustrating  $S_1$  and  $S_2$  exchange in addition to the activation intermediates.



**Extended Data Figure 5 | Comparison of two- and three-state models of  $^{19}\text{F}$ -labelled  $\text{A}_{2\text{A}}\text{R-V229C}$ .** **a**,  $^{19}\text{F}$  NMR  $T_2$  relaxation series of the  $^{19}\text{F}$ -labelled apo  $\text{A}_{2\text{A}}\text{R-V229C}$  receptor. **b**, Exponential fit to  $T_2$  for the downfield and upfield resonances, A and B in **a**. **c**, Deconvolution of the  $^{19}\text{F}$  NMR spectrum for  $^{19}\text{F}$ -labelled apo  $\text{A}_{2\text{A}}\text{R-V229C}$  receptor assuming a two-state model. The fitted line width of the upfield resonance is roughly twice that estimated from the  $T_2$  measurement, suggesting the upfield resonance may be better represented as a superposition of two Lorentzian lines, associated with  $S_3$  and  $S_{3'}$ , as discussed in the Supplementary Information. **d**, Spectral deconvolution of the  $^{19}\text{F}$  NMR spectrum of the  $^{19}\text{F}$ -labelled apo  $\text{A}_{2\text{A}}\text{R-V229C}$  receptor assuming three states. Note that

the most downfield peak is ascribed to  $S_{1-2}$ , which results from the rapid flickering of the ionic lock from 'on' ( $S_1$ ) to 'off' ( $S_2$ ), as evidenced by the CPMG measurements in Extended Data Fig. 4. Thus, we propose a total of four states, three of which may be spectroscopically resolved. The resonance frequencies chosen in the fit for  $S_3$  and  $S_{3'}$  were based on the observed peaks seen in the presence of agonists and those identified at pH 6, where  $S_3$  and  $S_{3'}$  are better resolved. The fitted line widths are also comparable to the homogeneous line widths, estimated from the above  $T_2$  experiment. Note that the difference spectrum (that is, the experimental spectrum minus spectral deconvolution) associated with the fit is shown in blue in **c** and **d**.

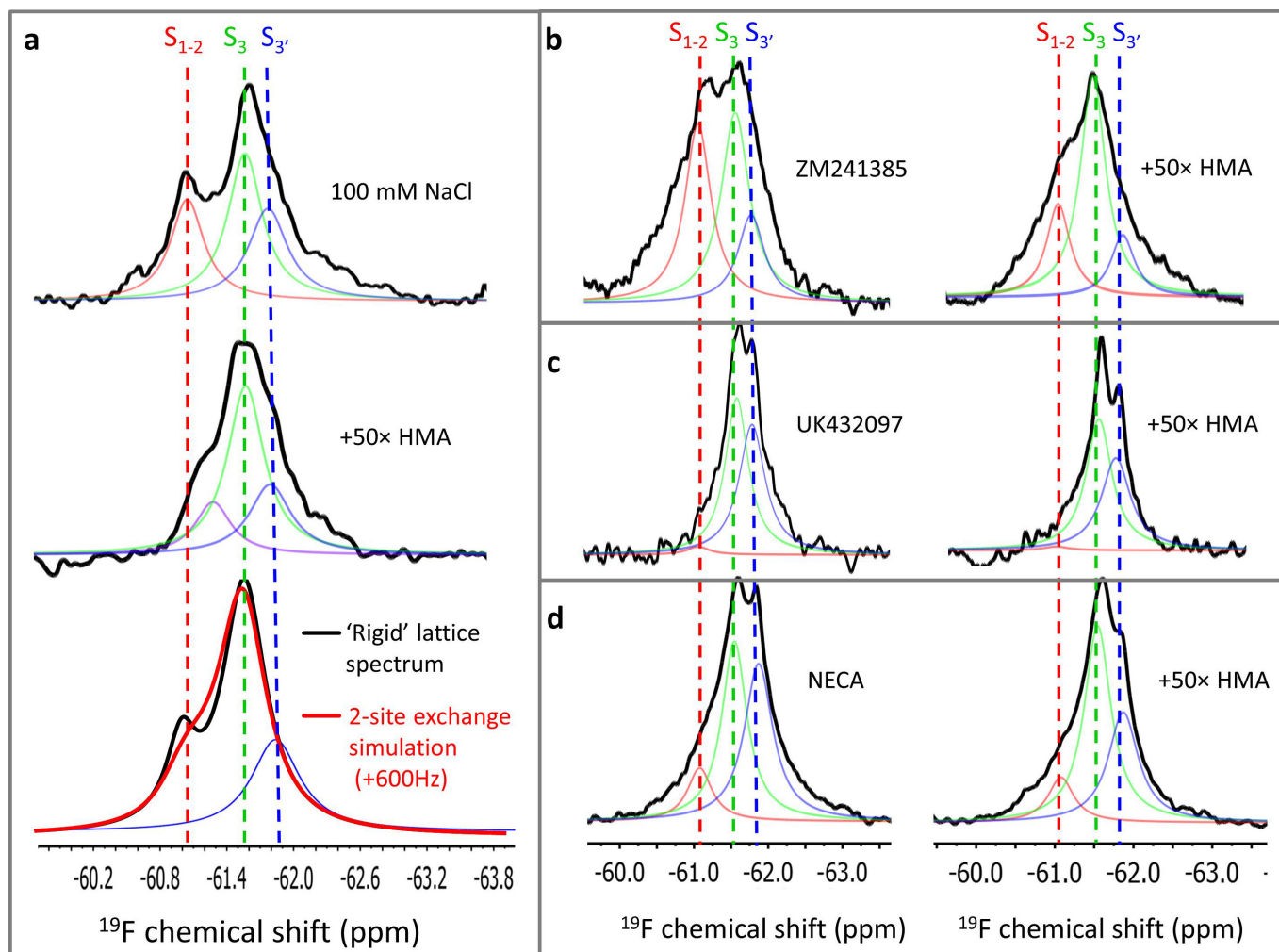




**Extended Data Figure 6 |  $^{19}\text{F}$  NMR spectra of  $^{19}\text{F}$ -labelled  $\text{A}_{2\text{A}}\text{R-V229C}$  in the presence of 50- or 100-fold excess of different ligands.**

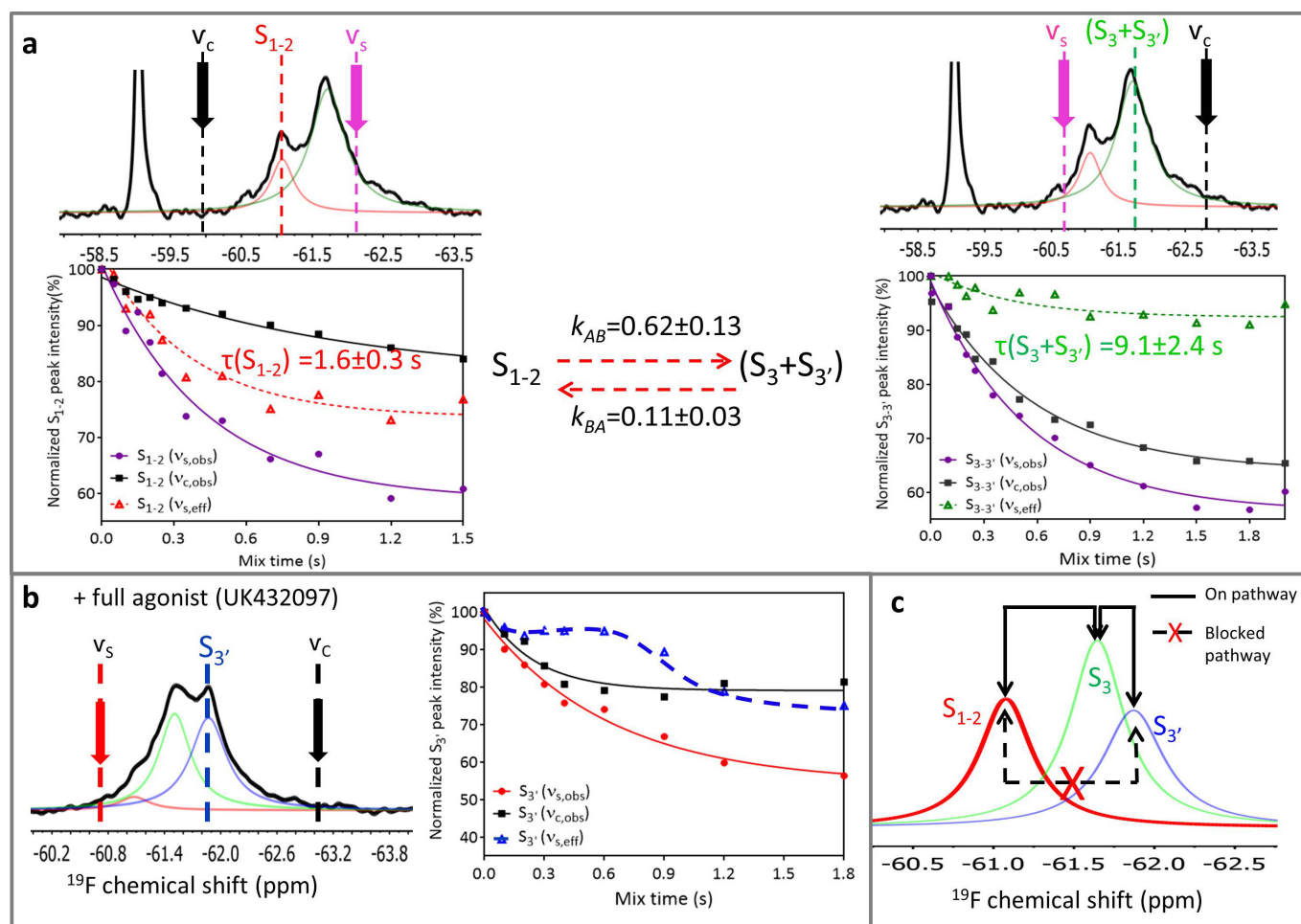
Representative ( $N=3$ )  $^{19}\text{F}$  NMR spectra as a function of ligands (inverse agonist (ZM241385), partial agonist (LUF5834), and full agonists UK432097, as shown in Fig. 1a. The downfield peak represents a reference peak resulting from the addition of  $10\text{ }\mu\text{M}$  bendroflumethazide. Note that a difference spectrum (shown in dark blue), corresponding to the

difference between the sum of the three deconvoluted resonances and the observed spectrum, is shown in each case. Note that the chemical shifts in the deconvolutions were referenced to the standard ( $-59.050\text{ ppm}$ ) and estimated to be  $-61.08\text{ ppm}$  ( $S_{1-2}$  (red)),  $-61.60\text{ ppm}$  ( $S_3$  (green)), and  $-61.85\text{ ppm}$  ( $S_{3'}$  (blue)), respectively. Corresponding line widths were estimated to be 220 Hz, 230 Hz, and 260 Hz, respectively.



**Extended Data Figure 7 | The role of HMA in the receptor activation process.** **a**,  $^{19}\text{F}$  NMR spectra of  $^{19}\text{F}$ -labelled apo  $\text{A}_{2\text{A}}\text{R-V229C}$  and  $^{19}\text{F}$ -labelled  $\text{A}_{2\text{A}}\text{R-V229C}$  in the presence of saturating amounts of the amiloride ligand 5-(*N,N*-hexamethylene) amiloride (HMA). Addition of 50-fold excess of HMA results in an increase in the  $S_3$  fraction and an apparent exchange broadening and slight coalescence of  $S_{1-2}$  and  $S_3$ , which are represented by the deconvolutions in lavender and green, respectively. After accounting for the exchange process between  $S_{1-2}$  and  $S_3$  by assuming  $k_{\text{ex}} = 600$  Hz, the simulated spectrum (shown in red) compares

favourably with the observed spectrum. If we assume that exchange between  $S_{1-2}$  is slow, we then obtain the 'rigid' lattice spectrum, shown in black. **b–d**,  $^{19}\text{F}$  NMR spectra of  $^{19}\text{F}$ -labelled  $\text{A}_{2\text{A}}\text{R-V229C}$  showing the effect of the addition of 50-fold excess of HMA to saturating amounts of inverse agonist (100  $\times$  ZM241385) and agonist (50  $\times$  UK432097 or 100  $\times$  NECA). In all cases, addition of HMA competes with the bound ligand and establishes a greater fraction of the  $S_3$  state. The three deconvolved resonances are shown in red, green, and blue.



**Extended Data Figure 8 | Saturation transfer experiments of  $^{19}\text{F}$ -labelled A<sub>2A</sub>R-V229C.** **a**,  $^{19}\text{F}$  NMR spectra of  $^{19}\text{F}$ -labelled apo A<sub>2A</sub>R-V229C with corresponding decay curves associated with continuous wave saturation of either the active state ensemble,  $S_3 + S_{3'}$ , or the inactive state ensemble,  $S_{1-2}$ , are provided in the left and right columns, respectively. To account for off-resonant saturation effects, a control experiment was performed at a frequency,  $\nu_c$ , such that the peak of interest was equidistant to the saturation frequency,  $\nu_s$ , and the control frequency,  $\nu_c$ . The response of the peak of interest (that is,  $S_{1-2}$  and  $S_3 + S_{3'}$  in the left and right panels, respectively) to saturation at the control frequency,  $\nu_c$ , is represented by black squares. Similarly, the response of the peak of interest to saturation at  $\nu_s$  is shown in violet while the effective responses, accounting for off-resonant saturation, are shown in red ( $S_{1-2}$ ) and green ( $S_3 + S_{3'}$ ).

On the basis of the effective decay profiles, and using a two-site exchange model, the lifetime of the inactive state ensemble and active states is estimated to be 1.6 s and 9 s. Spectral deconvolutions allow us to estimate the populations,  $p(S_{1-2})$  and  $p(S_3 + S_{3'})$ , to be 0.28 and 0.72, respectively. Using the fitted forward rate constant,  $k_{AB} = 0.62 \text{ s}^{-1}$ , the reverse rate constant is estimated to be  $k_{BA} = 0.24 \text{ s}^{-1}$ , assuming  $k_{AB} \times p(S_{1-2}) = k_{BA} \times p(S_3 + S_{3'})$ . In contrast, the response to the saturation of  $S_{1-2}$  provided an estimate of  $k_{BA} = 0.11 \pm 0.03 \text{ s}^{-1}$ . **b**, Saturation transfer experiments of full agonist UK432097-bound  $^{19}\text{F}$ -labelled A<sub>2A</sub>R-V229C. The effective decay curve (blue dashed line), associated with saturation of  $S_{1-2}$  is consistent with a process where  $S_{3'}$  magnetization is exchanged with  $S_{1-2}$  via  $S_3$ , as suggested by the figure in **c**. **c**, Model for presumed exchange pathway between  $S_{1-2}$ ,  $S_3$ , and  $S_{3'}$ .

Extended Data Table 1 | Primers/gene fragments used to construct plasmids for this study

Primer/ fragment	Sequences	Constructs
TEV fragment	5'-TCTAACAACAACAACAACAACAACAACAACAACCTTGGCGA AAACTTGTATTTCAGGGCGCT-3'	pPIC9K_Fa-Factor-Flag-TEV- A2aARTr316-H10
PicP1-J	5'-ATTCGAAGGATCCAAACGATGAGATTTC-3' (BamHI)	pPIC9K_Fa-Factor-Flag-TEV- A2aARTr316-H10
PicP2-J	5'-GTTGTTGTTGTTGTTAGACTTATCGTCATCGTCCTTGTAGTCTC-3'	pPIC9K_Fa-Factor-Flag-TEV- A2aARTr316-H10
PicP3-H	5'GGACGATGACGATAAGTCTAACAACAACAACAACAACAACAACA AC-3'	pPIC9K_Fa-Factor-Flag-TEV- A2aARTr316-H10
PicP41-H	5'- TGCCTTGAAAGGTTCTTGCTGCC-3'	pPIC9K_Fa-Factor-Flag-TEV- A2aARTr316-H10
PicP4-H	5'-ATTCGCGGCCGCTCAGTGATGGTGATGGTGATGGTGATGGTG ATGTGCCTTGAAAGGTTCTTGCTGCC-3' (NotI)	pPIC9K_Fa-Factor-Flag-TEV- A2aARTr316-H10
P <sub>A2a_V229C</sub>	5'- CCACACTGCAGAAGGAGTGCCATGCTGCCAAGTCAC-3'	pPIC9K_Fa-Factor-Flag- TEV-A2aARTr316-H10_V229C
PF <sub>AOX1</sub>	5'- GACTGGTTCCAATTGACAAGC-3'	sequencing
PR <sub>AOX1</sub>	5'- GGCAAATGGCATTCTGACATCCT-3'	sequencing



# Architecture of the mitochondrial calcium uniporter

Kirill Oxenoid<sup>1\*</sup>, Ying Dong<sup>2\*</sup>, Chan Cao<sup>1,3\*</sup>, Tanxing Cui<sup>1\*</sup>, Yasemin Sancak<sup>4</sup>, Andrew L. Markhard<sup>4</sup>, Zenon Grabarek<sup>4</sup>, Liangliang Kong<sup>2</sup>, Zhijun Liu<sup>2</sup>, Bo Ouyang<sup>2</sup>, Yao Cong<sup>2</sup>, Vamsi K. Mootha<sup>4</sup> & James J. Chou<sup>1,2</sup>

**Mitochondria from many eukaryotic clades take up large amounts of calcium ( $\text{Ca}^{2+}$ ) via an inner membrane transporter called the uniporter. Transport by the uniporter is membrane potential dependent and sensitive to ruthenium red or its derivative Ru360 (ref. 1). Electrophysiological studies have shown that the uniporter is an ion channel with remarkably high conductance and selectivity<sup>2</sup>.  $\text{Ca}^{2+}$  entry into mitochondria is also known to activate the tricarboxylic acid cycle and seems to be crucial for matching the production of ATP in mitochondria with its cytosolic demand<sup>3</sup>. Mitochondrial calcium uniporter (MCU) is the pore-forming and  $\text{Ca}^{2+}$ -conducting subunit of the uniporter holocomplex, but its primary sequence does not resemble any calcium channel studied to date. Here we report the structure of the pore domain of MCU from *Caenorhabditis elegans*, determined using nuclear magnetic resonance (NMR) and electron microscopy (EM). MCU is a homo-oligomer in which the second transmembrane helix forms a hydrophilic pore across the membrane. The channel assembly represents a new solution of ion channel architecture, and is stabilized by a coiled-coil motif protruding into the mitochondrial matrix. The critical DXXE motif forms the pore entrance, which features two carboxylate rings; based on the ring dimensions and functional mutagenesis, these rings appear to form the selectivity filter. To our knowledge, this is one of the largest membrane protein structures characterized by NMR, and provides a structural blueprint for understanding the function of this channel.**

Recently, genomic approaches have revealed the full molecular machinery of the uniporter holocomplex (uniuplex)<sup>4–8</sup>. In vertebrates, this complex consists of the transmembrane (TM) domain containing protein MCU, its inactive paralogue MCUb, and an accessory single-pass TM peptide called EMRE. In addition, the complex includes two paralogous, EF-hand  $\text{Ca}^{2+}$ -binding proteins MICU1 and MICU2 in the intermembrane space. Current models of the uniporter indicate that MCU is the pore-forming subunit, and that MICU1/2 are  $\text{Ca}^{2+}$ -sensing proteins that gate the activity of the pore based on cytosolic  $\text{Ca}^{2+}$  concentrations<sup>9</sup>. EMRE is metazoan specific and appears to have two key functions: it maintains the pore in an open conformation, while additionally transducing MICU1/2  $\text{Ca}^{2+}$  sensing to the pore<sup>7</sup>.

There is consensus now based on several lines of evidence that MCU encodes the pore-forming subunit. First, loss of MCU leads to complete abrogation of uniporter current<sup>4,10</sup>. Second, expression of the MCU orthologue from *Dictyostelium*, an organism that does not have EMRE, is alone sufficient to reconstitute uniporter activity in yeast<sup>11</sup>. Third, MCU has conserved acidic residues in the DXXE sequence motif at the putative entrance of the channel that are essential for  $\text{Ca}^{2+}$  uptake<sup>4,6</sup>. Fourth, MCU monomers oligomerize into a higher molecular mass assembly<sup>4,11</sup>, as would be required of a pore. Fifth, a point mutation in MCU confers resistance to Ru360 while preserving

uniporter current<sup>4,10</sup>, providing compelling biochemical evidence that MCU is the probable mechanistic target of Ru360. At present, information on the architecture of this pore protein is lacking. For example, we do not know what its oligomeric state is, which TM helix forms the pore, or what the structural basis of channel regulation is.

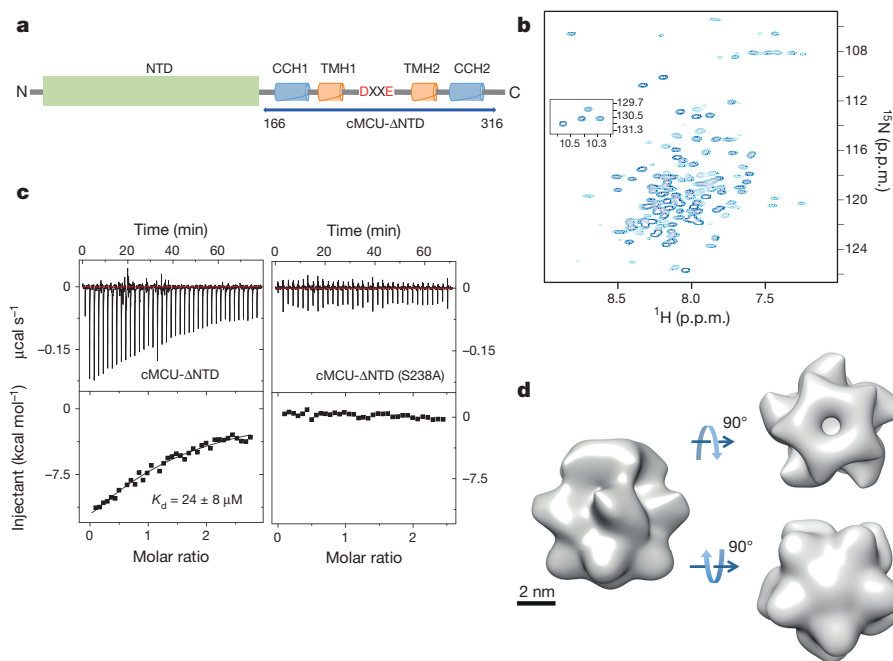
To determine the structure of the pore domain of MCU, we used an approach that combines EM and NMR. MCU is predicted to be at least a tetramer<sup>8</sup>. Thus the full-length complex (>160 kDa) is too large for *de novo* structure determination by present solution NMR technology. The protein contains a soluble amino-terminal domain (NTD; ~165 residues) that may be dispensable for channel activity<sup>12</sup> (Fig. 1a). We screened several constructs of MCU with a deleted NTD, and found that the one from *C. elegans* (cMCU- $\Delta$ NTD) (Extended Data Fig. 1) could be expressed to high levels in *Escherichia coli*. The protein was extracted using foscholine-14 detergent followed by ion-exchange and size-exclusion chromatography in physiological buffer at pH 6.5 (see Methods). The purified cMCU- $\Delta$ NTD in foscholine-14 formed pentamers as suggested by size-exclusion chromatography coupled to multi-angle light scattering (SEC-MALS) and crosslinking (Extended Data Fig. 2), and generated NMR spectra of good chemical shift dispersion and resolution (Fig. 1b and Extended Data Fig. 3). Under the sample condition, cMCU- $\Delta$ NTD bound Ru360, and introduction of the Ser238Ala mutation, which was shown previously to confer resistance<sup>4,10</sup>, reduced the inhibitor binding (Fig. 1c).

We first performed negative-stain EM reconstruction of the cMCU- $\Delta$ NTD oligomers, after diluting NMR sample, using the single particle analysis method (see Methods), with the goal of obtaining a global structural framework to aid subsequent structure determination by NMR. From the 12,860 automatically picked particles, we obtained a reconstructed 3D density map refined to a resolution of ~18 Å (Extended Data Fig. 4). The EM map has a roughly cylindrical shape with five-fold symmetry (Fig. 1d), indicating that cMCU- $\Delta$ NTD forms pentamers. Despite the low resolution, the map showed many interesting features. One end of the cylinder has a deep hole, probably corresponding to the TM pore domain, whereas the opposite end is solid, possibly due to the formation of a coiled-coil (CC) complex by the predicted CC helix at the carboxy terminus (Fig. 1a,d). Moreover, each subunit appears to have three levels, and the middle level exhibits an unusual bulge.

The pentameric complex formed by cMCU- $\Delta$ NTD has a total molecular mass of ~90,375 Da and thus represents a formidable challenge to NMR spectroscopy. Further challenge came from poor sample stability at temperatures >23 °C. Consequently, we recorded NMR data at two different temperatures: 23 °C for collecting the bulk structural restraints and 33 °C for providing information on protein regions that showed weak NMR signals at 23 °C. The general approach we used for structure determination involves: (i) determination of

<sup>1</sup>Department of Biological Chemistry and Molecular Pharmacology, Harvard Medical School, Boston, Massachusetts 02115, USA. <sup>2</sup>State Key Laboratory of Molecular Biology, National Center for Protein Science Shanghai, Shanghai Institute of Biochemistry and Cell Biology, Shanghai Science Research Center, Chinese Academy of Sciences, Shanghai 200031, China. <sup>3</sup>State Key Laboratory of Elemento-Organic Chemistry and College of Chemistry, Nankai University, Tianjin 300071, China. <sup>4</sup>Department of Molecular Biology and Howard Hughes Medical Institute, Massachusetts General Hospital, Boston, Massachusetts 02114, USA.

\*These authors contributed equally to this work.



**Figure 1 | NMR and EM characterization of cMCU-ΔNTD.**

**a**, Domain organization of MCU (predicted using programs COILS<sup>25</sup> and TMHMM<sup>26</sup>). **b**,  $^1\text{H}$ - $^{15}\text{N}$  TROSY heteronuclear single quantum coherence (HSQC) spectrum of ( $^{15}\text{N}$ ,  $^2\text{H}$ )-labelled cMCU-ΔNTD oligomer reconstituted in foscholine-14, recorded at 900 MHz and 23 °C. The peaks in the inset correspond to tryptophan side-chain amines. **c**, Isothermal titration calorimetry (ITC) analysis of Ru360 binding to cMCU-ΔNTD (left) and the Ser238Ala mutant (right) under the NMR sample condition.

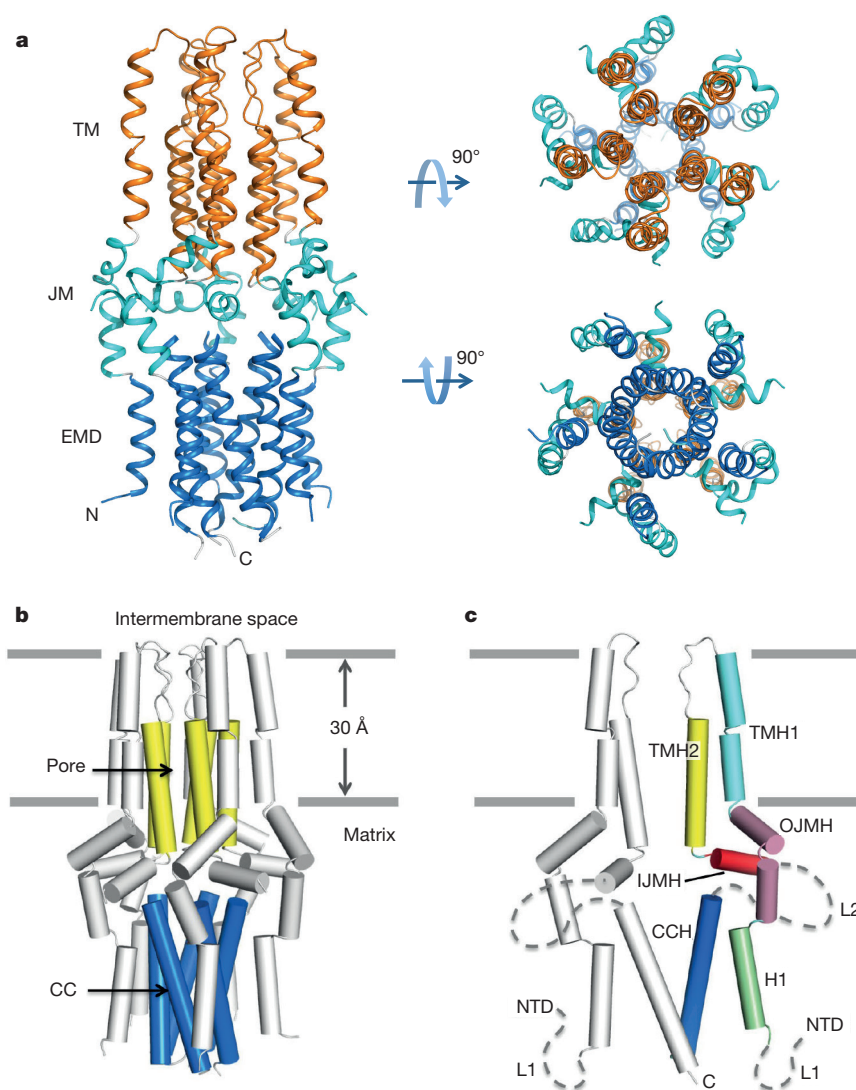
The top and bottom graphs show the raw data ( $\mu\text{cal s}^{-1}$  versus time) and normalized integration data ( $\text{kcal mol}^{-1}$  of injectant versus molar ratio; molar ratio = Ru360:cMCU-ΔNTD pentamer), respectively. Data fitting yields dissociation constant ( $K_d$ ) =  $24 \pm 8 \mu\text{M}$ . **d**, Negative stain EM reconstruction of the cMCU-ΔNTD oligomers using a protein sample prepared in the same way as NMR samples. Views of the final 3D volumes of cMCU-ΔNTD oligomer filtered at 18 Å.

local structures of the monomers, and (ii) assembly of the oligomer with intermonomer distance restraints<sup>13–15</sup>. The secondary structures of the monomers in the oligomeric complex were determined mainly using local distance restraints derived from nuclear Overhauser enhancements (NOEs). We then used a mixed sample with differentially labelled subunits to measure exclusively NOEs between the  $^{15}\text{N}$ -attached protons of one subunit and non-exchangeable protons of the neighbouring subunits. This experiment provided key intermonomer NOEs defining the CC, the TM pore, and the ion-selectivity domains (Extended Data Fig. 5). The initial structural solution then allowed iterative assignment of additional intermonomer and long-range NOEs using standard NOE experiments. The structure was determined using 2,070 local and 150 long-range intramonomer distance restraints and 220 intermonomer restraints (Extended Data Fig. 6a and Extended Data Table 1).

The NMR structure of cMCU-ΔNTD shows a well-packed pentamer with an overall cylindrical shape similar to the EM map (Extended Data Fig. 6b–e). The EM structure is slightly shorter and wider than the NMR structure, which could have been caused by specimen flattening, a phenomenon commonly observed in negative stain EM<sup>16</sup>. Both have a star-like appearance when viewed from top or bottom, and feature five bulges in the middle of the assembly that curve in the same direction (Figs 2a and 1d). The more detailed NMR structure revealed the architecture of the channel assembly. The inner core of the pentamer is formed with the second TM helix (TMH2; 244–260) and the coiled-coil helix (CCH; residues 293–316) (Fig. 2b, c). While the TMH2s pack into a five-helix bundle having a largely polar pore across the membrane, the CCH outside the membrane forms a CC pentamer with a hydrophobic core, which may contribute to stabilizing the TM pore structure. As an independent validation of the pentameric assembly of the CCH, we showed by chemical crosslinking that a peptide containing the predicted CC domain (residues 289–316) in water forms a pentamer in agreement with the oligomeric state of cMCU-ΔNTD (Methods; Extended Data Fig. 7). The two core domains are

structurally supported by peripheral helices: the TM pore domain is wrapped by the first TM helix (TMH1; residues 215–234) through contacts between TMH1 and TMH2, and the extramembrane CC domain is wrapped by the first helix of cMCU-ΔNTD (H1; residues 180–193) (Fig. 2b, c). The two core domains are not continuous as TMH2 ends at Tyr260, which is immediately followed by the inner juxtamembrane helix (IJMH; residues 262–271) that orients at a wide angle relative to TMH2. The IJMH turns into an unstructured loop (L2; residues 272–292) before the beginning of the CCH. NMR peaks for many of the residues in L2 were not observed, possibly owing to solvent exchange and conformational heterogeneity. The two core domains appear to be held together on the periphery by the outer juxtamembrane helix (OJMH; residues 195–213), which contains a kink around Glu204 and interacts with the IJMH. Despite the apparent structure, NMR signals for IJMH and OJMH are mostly weak due to exchange broadening, suggesting the membrane proximal region of cMCU-ΔNTD is intrinsically unstable. In addition to L2, another unstructured region is the loop preceding H1 (L1; residues 166–179); it is unstructured due to the absence of NTD as residues 166–171 in the corresponding NTD of human MCU form a short helix<sup>12</sup>.

The structure suggests a  $\text{Ca}^{2+}$  flow pathway through the MCU pore. The critical DXXE motif connecting TMH1 and TMH2 forms a pentameric barrel that appears to be the mouth of the pore (Fig. 3a). Inside the barrel, both acidic residues are in position to form two carboxylate rings. The Asp240 ring is solvent exposed, and the Glu243 ring is located deeper, guarding the entrance of the TMH2 pore commencing at Pro244. The precise ring sizes cannot be measured in our structure because the Asp240 and Glu243 side chains could not be precisely defined using the available NOE restraints (Fig. 3b), but the diameter of the ring defined by C $\beta$  is 7 Å and 11 Å for Asp240 and Glu243, respectively. These values are smaller on average than those of the asparagine ring in the CorA  $\text{Mg}^{2+}$  channel (11 Å)<sup>17–20</sup> or the glutamate ring in the Orai  $\text{Ca}^{2+}$  channel (12 Å)<sup>21</sup>, suggesting that  $\text{Ca}^{2+}$  should be partially dehydrated by the Asp240 and Glu243 rings.



**Figure 2 | Structure of the cMCU- $\Delta$ NTD pentamer.** **a**, Ribbon representation of the cMCU- $\Delta$ NTD structure displaying three distinct layers that correspond to the transmembrane (TM; orange), juxtamembrane (JM; cyan), and extramembrane domain (EMD; blue) regions, respectively. **b**, Cartoon representation of the cMCU- $\Delta$ NTD pentamer showing the formation of the uniporter core, which consists of the TM pore formed by TMH2 (yellow) and the coiled-coil pentamer

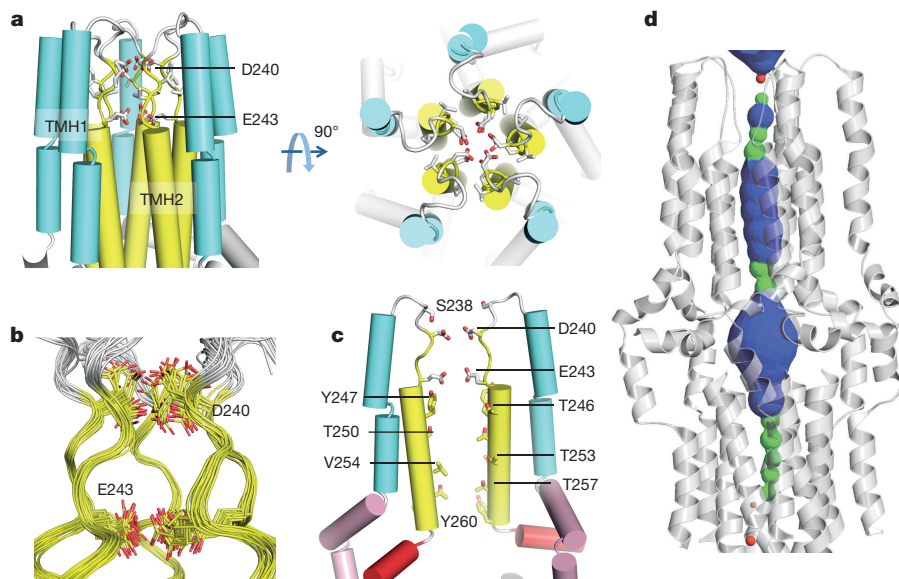
formed by CCH (marine). The structure is placed in the presumed membrane such that the peripheral hydrophobic residues in the TM domain are lipid facing. **c**, Cartoon representation of two subunits of the cMCU- $\Delta$ NTD pentamer showing the folding of individual subunits. The helical segments are defined in the text. The dashed lines labelled as L1 and L2 correspond to the unstructured regions of the protein.

Passing the acidic rings,  $\text{Ca}^{2+}$  can readily move through the TM pore, which is hydrophilic and lined mostly with threonines (Fig. 3c and Extended Data Fig. 8). The TM pore ends with Tyr260. Upon exiting the TM pore from the C-terminal end,  $\text{Ca}^{2+}$  will be in a very hydrophilic chamber enclosed by IJMH and OJMH on the side and the CC domain on the bottom (Fig. 2b, c).  $\text{Ca}^{2+}$  cannot exit through the CC domain because this domain has a solid hydrophobic core. Surface plot also shows that interaction between IJMH and OJMH seals the side of the chamber leaving no visible holes for ions to exit laterally (Fig. 3d). Thus, the structure represents the closed conformation of the channel, which is consistent with the fact that metazoan MCU in the absence of EMRE does not transport  $\text{Ca}^{2+}$  (refs 7, 11). We speculate that conformational rearrangements induced by EMRE binding are needed to allow  $\text{Ca}^{2+}$  to exit. By nature of being unstable, the IJMH, OJMH and L2 segments of each MCU monomer are the likely regions to undergo conformational changes that would create lateral exit paths near the membrane, each activated by its own EMRE peptide.

The structure of cMCU- $\Delta$ NTD provides a framework for better understanding human MCU (HsMCU), especially because HsMCU

and cMCU are orthologous, with greater than 40% sequence identity, sharing the same protein domain organization<sup>22</sup>. We previously reported the use of HEK-293T cells in which we knocked out HsMCU<sup>7</sup>. Such cells do not exhibit mitochondrial  $\text{Ca}^{2+}$  uptake, and represent an excellent system into which we can introduce mutant alleles to evaluate their effect on mitochondrial  $\text{Ca}^{2+}$  uptake. We created the NTD deletion in HsMCU (HsMCU- $\Delta$ NTD), corresponding to deletion in cMCU- $\Delta$ NTD, and found that the mutant is able to complement the need for HsMCU, demonstrating that the NTD in the human protein is dispensable for calcium uptake activity (Extended Data Fig. 9). We then created a cysteine-free mutant of HsMCU (HsMCU<sup>CF</sup>) akin to cMCU- $\Delta$ NTD, and found that it could rescue  $\text{Ca}^{2+}$  transport in human cells lacking MCU (Fig. 4a). We next proceeded to introduce mutations into HsMCU<sup>CF</sup> to validate our NMR structure of cMCU- $\Delta$ NTD. Glu257 (Glu236 in cMCU) is highly conserved across eukaryotic MCU orthologues, but it is not conserved in MCUb, a non-conducting human paralogue of MCU. It was proposed that this residue is key to  $\text{Ca}^{2+}$  conductance<sup>4,8</sup> because transient expression of Glu257Ala mutant on a partial MCU knockdown background failed





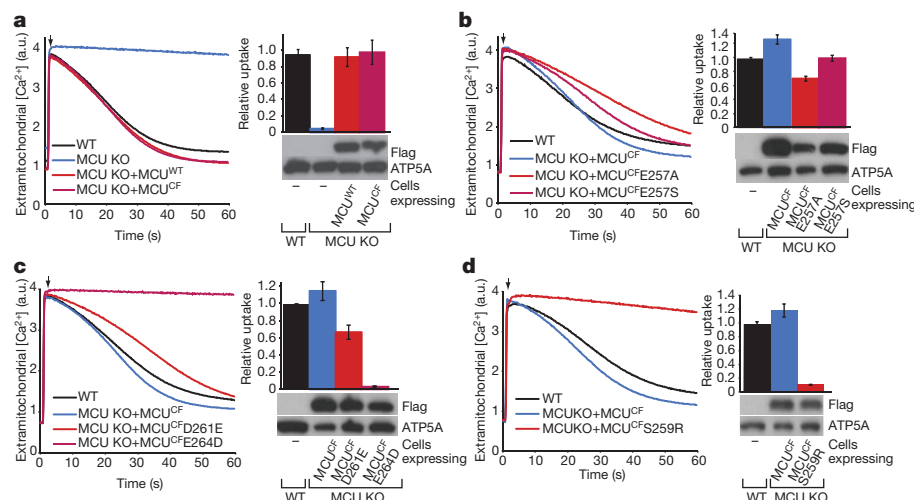
**Figure 3 | Architecture of the pore and ion selectivity filter.** **a**, Cartoon representation of the TM domain displaying the mini barrel at the mouth of the TM pore that contains the DXXE  $\text{Ca}^{2+}$  selectivity elements. **b**, Enlarged view of the DXXE-containing region for showing side-chain conformational diversity of Asp240 and Glu243. **c**, TM domains of

subunits 1 and 3 for showing the pore-lining residues. **d**, The pore surface calculated using the program HOLE<sup>27</sup>. The region of the channel colored in green is only wide enough to allow passage of one water molecule, whereas the blue portion can accommodate two or more water molecules. The red region is too narrow to allow any water to pass through.

to rescue calcium transport fully. However, our structure (Fig. 3a, c) places this residue outside the entrance of the pore, and predicts it to be dispensable for  $\text{Ca}^{2+}$  uptake. Indeed, consistent with this prediction, we find that when Glu257Ala or Glu257Ser mutants are stably expressed at levels comparable to that of wild-type protein on a clean MCU knockout background, they conduct  $\text{Ca}^{2+}$  (Fig. 4b). In contrast to Glu257, substitutions of Asp or Glu within the DXXE motif with Ala disrupted mitochondrial  $\text{Ca}^{2+}$  uptake<sup>4</sup>. Our structure shows that the side chains of these residues provide two carboxylate rings that might be involved in  $\text{Ca}^{2+}$  permeation and/or selectivity. Given how critical these two residues appear to be in the structure, we made very conservative mutations, namely Asp261Glu and Glu264Asp (Asp240 and Glu243 in cMCU). While Asp261Glu is partially functional, we note that Glu264Asp has completely lost its ability to conduct  $\text{Ca}^{2+}$  despite

expressing properly (Fig. 4c), underscoring the critical importance of this residue for permeation.

We next considered the mechanism of action of the drug Ru360, a well-known inhibitor of the uniporter. Ru360 inhibits the human uniporter with nanomolar potency in isolated mitochondria or in permeabilized cells. We previously identified Ser259 (Ser238 in cMCU) as being potentially important for the mechanism of action of Ru360, since the Ser259Ala mutant permits  $\text{Ca}^{2+}$  conductance but confers nearly complete resistance against Ru360 (refs 4, 10). In our structure of cMCU- $\Delta$ NTD, this serine residue lies at the apex of the pore, before the mini-barrel, raising the hypothesis that Ru360 operates by simply obstructing the pore. If this hypothesis is correct, then the introduction of bulky residues into this position ought to occlude the pore. In fact, Ser259Arg is well expressed and does not permit any  $\text{Ca}^{2+}$  to



**Figure 4 | Functional mutagenesis of HsMCU inspired by the cMCU- $\Delta$ NTD structure.** **a**, Cysteine-free MCU ( $\text{MCU}^{\text{CF}}$ ) rescues mitochondrial calcium uptake to the same extent as wild-type MCU ( $\text{MCU}^{\text{WT}}$ ) in MCU knockout (KO) cells. **b**, Mutation of MCU Glu257 to Ala or Ser does not impair its function. **c**, Mutation of Asp261 to Glu permits ion permeation through MCU whereas Glu264 to Asp impairs function. **d**, Mutation of

Ser259 to Arg impairs MCU function. **a–d**, Representative traces of  $\text{Ca}^{2+}$  uptake in digitonin-permeabilized cells after addition of  $50 \mu\text{M}$   $\text{CaCl}_2$  are shown on left. The bar graph shows the rate of  $\text{Ca}^{2+}$  uptake relative to wild-type HEK-293T cells (mean  $\pm$  s.d.,  $n = 4$ ). Cell lysates were analysed by immunoblotting using anti-Flag antibody to detect expression of MCU protein. ATP5A was used as loading control.



enter, probably mimicking the effect of Ru360 (Fig. 4d). Collectively, these experiments demonstrate the value of our structure for understanding the function of human MCU.

The cMCU- $\Delta$ NTD structure represents a new solution of ion channel architecture. The bacterial and mitochondrial  $\text{Mg}^{2+}$  channel, CorA/MRS2, also forms a pentamer and has a similar domain organization based on primary and secondary structures, but it has very different tertiary and quaternary fold<sup>19,20</sup>. The TM domains of both CorA and MCU are composed of ten helices (two from each subunit) arranged in two concentric layers. Whereas the CorA pore is formed by TMH1, the MCU pore is formed by TMH2. The selectivity filter in MCU appears to be composed of two carboxylate rings: the solvent-accessible aspartate and the internal glutamate rings (Fig. 3a, b). In CorA, the solvent-accessible ring is made of five asparagines, and the slightly larger internal ring of backbone carbonyls of glycine residues<sup>17</sup>. By contrast, the  $\text{Ca}^{2+}$  release activated  $\text{Ca}^{2+}$  channel Orai has only one ring of six glutamate side chains in the filter region<sup>21</sup>. Another major difference between CorA and Orai on the one hand and MCU on the other is that after the selectivity filter, the pore in both CorA and Orai becomes hydrophobic for several helical turns. In MCU, however, the pore remains hydrophilic suggesting an explanation for fast ion conduction by this channel. Finally, in Orai and CorA, the ion pathway continues far beyond the membrane bilayer, but in MCU the pore ends after traversing the bilayer, with ions likely exiting laterally near the membrane. In this respect the architecture of cMCU- $\Delta$ NTD is reminiscent of some channel proteins that belong to the diverse family of pentameric ligand gated ion channels (pLGIC), which includes serotonin 5-HT<sub>3</sub> and nicotinic acetylcholine receptors<sup>23,24</sup>.

In conclusion, the combined use of NMR and EM enabled the characterization of the overall architecture of the MCU, which is also one of the largest membrane protein complexes studied by NMR. The structure represents a new pore architecture for conducting  $\text{Ca}^{2+}$  as its ion-selectivity filter, ion-conducting pore, and the arrangement of the extramembrane domains are all different from the known  $\text{Ca}^{2+}$  channel structures. Although the reported structure does not show  $\text{Ca}^{2+}$  exit, which is consistent with being non-conducting in the absence of EMRE, the overall architecture suggests lateral  $\text{Ca}^{2+}$  exit near the middle of the channel complex. Several outstanding questions remain. The structure shows the selectivity filter and the pore, yet the detailed mechanism of  $\text{Ca}^{2+}$  binding, transport, and exit is unknown, including the function of the L2 loop that could block the exit, but is unstructured in our model. Although we have provided compelling evidence that cMCU forms a pentamer *in vitro*, the oligomeric state *in vivo* remains to be confirmed. Finally, MCU interactions with its TM partner EMRE and its regulators MICU1/2 are important subjects of future investigation.

**Online Content** Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

**Received 22 October 2015; accepted 8 March 2016.**

**Published online 2 May 2016.**

- Gunter, T. E. & Pfeiffer, D. R. Mechanisms by which mitochondria transport calcium. *Am. J. Physiol.* **258**, C755–C786 (1990).
- Kirichok, Y., Krapivinsky, G. & Clapham, D. E. The mitochondrial calcium uniporter is a highly selective ion channel. *Nature* **427**, 360–364 (2004).
- Denton, R. M. & McCormack, J. G. The role of calcium in the regulation of mitochondrial metabolism. *Biochem. Soc. Trans.* **8**, 266–268 (1980).
- Baughman, J. M. *et al.* Integrative genomics identifies MCU as an essential component of the mitochondrial calcium uniporter. *Nature* **476**, 341–345 (2011).
- Perocchi, F. *et al.* MICU1 encodes a mitochondrial EF hand protein required for  $\text{Ca}^{2+}$  uptake. *Nature* **467**, 291–296 (2010).

- De Stefani, D., Raffaello, A., Teardo, E., Szabo, I. & Rizzuto, R. A forty-kilodalton protein of the inner membrane is the mitochondrial calcium uniporter. *Nature* **476**, 336–340 (2011).
- Sancak, Y. *et al.* EMRE is an essential component of the mitochondrial calcium uniporter complex. *Science* **342**, 1379–1382 (2013).
- Raffaello, A. *et al.* The mitochondrial calcium uniporter is a multimer that can include a dominant-negative pore-forming subunit. *EMBO J.* **32**, 2362–2376 (2013).
- Kamer, K. J. & Mootha, V. K. The molecular era of the mitochondrial calcium uniporter. *Nature Rev. Mol. Cell Biol.* **16**, 545–553 (2015).
- Chaudhuri, D., Sancak, Y., Mootha, V. K. & Clapham, D. E. MCU encodes the pore conducting mitochondrial calcium currents. *eLife* **2**, e00704 (2013).
- Kovács-Bogdán, E. *et al.* Reconstitution of the mitochondrial calcium uniporter in yeast. *Proc. Natl Acad. Sci. USA* **111**, 8985–8990 (2014).
- Lee, Y. *et al.* Structure and function of the N-terminal domain of the human mitochondrial calcium uniporter. *EMBO Rep.* (2015).
- Van Horn, W. D. *et al.* Solution nuclear magnetic resonance structure of membrane-integral diacylglycerol kinase. *Science* **324**, 1726–1729 (2009).
- Schnell, J. R. & Chou, J. J. Structure and mechanism of the M2 proton channel of influenza A virus. *Nature* **451**, 591–595 (2008).
- QuYang, B. *et al.* Unusual architecture of the p7 channel from hepatitis C virus. *Nature* **498**, 521–525 (2013).
- Cheng, Y. *et al.* Single particle reconstructions of the transferrin-transferrin receptor complex obtained with different specimen preparation techniques. *J. Mol. Biol.* **355**, 1048–1065 (2006).
- Guskov, A. *et al.* Structural insights into the mechanisms of  $\text{Mg}^{2+}$  uptake, transport, and gating by CorA. *Proc. Natl Acad. Sci. USA* **109**, 18459–18464 (2012).
- Pföh, R. *et al.* Structural asymmetry in the magnesium channel CorA points to sequential allosteric regulation. *Proc. Natl Acad. Sci. USA* **109**, 18809–18814 (2012).
- Eshaghi, S. *et al.* Crystal structure of a divalent metal ion transporter CorA at 2.9 angstrom resolution. *Science* **313**, 354–357 (2006).
- Lunin, V. V. *et al.* Crystal structure of the CorA  $\text{Mg}^{2+}$  transporter. *Nature* **440**, 833–837 (2006).
- Hou, X., Pedi, L., Diver, M. M. & Long, S. B. Crystal structure of the calcium release-activated calcium channel Orai. *Science* **338**, 1308–1313 (2012).
- Bick, A. G., Calvo, S. E. & Mootha, V. K. Evolutionary diversity of the mitochondrial calcium uniporter. *Science* **336**, 886 (2012).
- Hassaine, G. *et al.* X-ray structure of the mouse serotonin 5-HT<sub>3</sub> receptor. *Nature* **512**, 276–281 (2014).
- Unwin, N. Refined structure of the nicotinic acetylcholine receptor at 4 Å resolution. *J. Mol. Biol.* **346**, 967–989 (2005).
- Lupas, A., Van Dyke, M. & Stock, J. Predicting coiled coils from protein sequences. *Science* **252**, 1162–1164 (1991).
- Sonnhammer, E. L., von Heijne, G. & Krogh, A. A hidden Markov model for predicting transmembrane helices in protein sequences. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **6**, 175–182 (1998).
- Smart, O. S., Neduvellil, J. G., Wang, X., Wallace, B. A. & Sansom, M. S. HOLE: a program for the analysis of the pore dimensions of ion channel structural models. *J. Mol. Graph.* **14**, 354–360 (1996).

**Acknowledgements** We thank Y. Balazs for helping with ITC measurement and data analysis and Y. Zhang and M. Cao from the EM facility of NCPSS for their assistance with EM data collection. The NMR data were collected at the NMR facility of NCPSS and MIT-Harvard CMR (supported by NIH grant P41 EB-002026). This work was supported by CAS grant XDB08030301 and NIH grant GM094608 to J.J.C. V.K.M. is an Investigator of the Howard Hughes Medical Institute. Y.C. is supported by CAS grant XDB08030201. C.C. is supported by the China Scholarship Council.

**Author Contributions** T.C., Y.D., Y.S., C.C., V.K.M. and J.J.C. conceived the study; T.C., Y.S. and C.C. designed protein constructs for structural studies; C.C. and K.O. performed inhibitor binding studies; Y.S., A.L.M. and Z.G. performed structure guided functional experiments and analysis; Y.D., L.K. and Y.C. prepared EM samples and performed EM analysis. K.O., T.C., C.C. and J.J.C. collected NMR data and solved the structure; V.K.M., J.J.C. and K.O. wrote the paper and all authors contributed to editing of the manuscript.

**Author Information** The atomic structure coordinate and structural constraints are deposited in the Protein Data Bank (PDB) under the accession number 5ID3. The chemical shift values are deposited in the Biological Magnetic Resonance Data Bank (BMRB) under the accession number 30021. Reprints and permissions information is available at [www.nature.com/reprints](http://www.nature.com/reprints). The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to J.J.C. ([james\\_chou@hms.harvard.edu](mailto:james_chou@hms.harvard.edu)).

## METHODS

No statistical methods were used to predetermine sample size. The experiments were not randomized and the investigators were not blinded to allocation during experiments and outcome assessment.

**Protein sample preparation.** *E. coli*-codon optimized DNA encoding residues 167–316 of the *C. elegans* MCU (cMCU- $\Delta$ NTD) with C-terminal 6 $\times$ His tag was synthesized (GenScript) and cloned into the pET21a vector (Extended Data Fig. 1). BL21 (DE3) cells were transformed with the vector for expression. The cells were grown in either LB broth or isotopically labelled M9 minimal media at 37 °C until absorbance at 600 nm ( $A_{600\text{ nm}}$ ) reached 0.6–0.7. After induction with 0.2 mM isopropyl- $\beta$ -D-thiogalactopyranoside (IPTG), protein was expressed for 20 h at 18 °C. Cells were collected by centrifugation and resuspended and lysed by sonication in buffer A (20 mM HEPES, pH 7.4, 150 mM NaCl, 10  $\mu$ g ml<sup>-1</sup> lysozyme, 10  $\mu$ g ml<sup>-1</sup> DNase, 1 mM PMSF and 2 mM EDTA). After lysis, the inclusion bodies and membranes were collected by centrifugation at 39,000g for 1 h and resuspended in buffer B (20 mM HEPES, pH 7.4, 150 mM NaCl, 40 mM foscholine-14 and 2 mM EDTA), followed by stirring at 4 °C overnight. Insoluble aggregates were removed by centrifugation at 39,000g for 40 min. The supernatant was then subjected to Ni-NTA purification in buffer B with foscholine-14 concentration adjusted to 0.48 mM (4 $\times$  detergent CMC). The protein was eluted using 400 mM imidazole in the same buffer and dialysed against buffer C (25 mM CHES, pH 9.2, 50 mM NaCl and 2 mM EDTA) to remove imidazole. After dialysis, the sample was subjected to ion-exchange purification using the HiTrap Q HP (1 ml) column (GE Healthcare) in buffer C with added 0.48 mM foscholine-14. The protein was eluted using a NaCl gradient (0.05–1.00 M in 20 ml). Finally, the protein was further separated by size exclusion using the Superdex 200 10/300 GL column (GE Healthcare) in buffer D (20 mM MES, pH 6.4, 75 mM NaCl, 0.48 mM foscholine-14, 0.3 mM Na<sub>3</sub>N, and 2 mM EDTA). The elution peak fractions were analysed by SDS–PAGE (Extended Data Fig. 2) and pooled and concentrated to achieve 0.8 mM cMCU- $\Delta$ NTD (monomer) for NMR measurements. The final NMR sample buffer contains 20 mM MES, pH 6.4, 75 mM NaCl, ~27 mM foscholine-14, 0.3 mM Na<sub>3</sub>N, 2 mM EDTA and 5% D<sub>2</sub>O. Typical final yields of cMCU- $\Delta$ NTD were 2–3 mg of protein from 1 l of cell culture.

**SEC-MALS analysis of the cMCU- $\Delta$ NTD molecular mass.** The instrument setup used for the SEC-MALS experiment consisted of an Agilent 1260 Infinity Isocratic Liquid Chromatography System connected in series with a Wyatt Dawn Heleos II Multi-Angle Light Scattering (MALS) detector (Wyatt Technology) and a Wyatt Optilab T-rEX Refractive Index Detector (Wyatt Technology). Analytical size-exclusion chromatography was performed at room temperature using Superdex 200 10/300 GL column (GE Healthcare) equilibrated with a mobile phase containing 20 mM MES, pH 6.4, 75 mM NaCl, 0.48 mM foscholine-14, 0.3 mM Na<sub>3</sub>N and 2 mM EDTA. 100  $\mu$ l purified cMCU- $\Delta$ NTD sample at 4.0 mg ml<sup>-1</sup> was injected into the column and eluted at a flow rate of 0.4 ml min<sup>-1</sup>. The column effluent was monitored in-line with three detectors that simultaneously monitored UV absorption, light scattering and refractive index. The data from the three detectors were imported by the ASTRA software package, and the three-detector method<sup>28</sup> was used to determine the molecular mass.

**Characterization of the cMCU- $\Delta$ NTD oligomeric state by crosslinking.** The oligomeric state of cMCU- $\Delta$ NTD in the NMR sample was examined by chemical crosslinking using DTSSP (3,3'-dithiobis(sulfosuccinimidyl propionate) (Thermo Scientific)). The cMCU- $\Delta$ NTD sample was first dialysed overnight against the crosslinking reaction buffer (25 mM sodium phosphate, pH 7.5, and 50 mM NaCl). A stock solution of 50 mM DTSSP in the reaction buffer was prepared for addition to the protein solution. Five samples each containing 0.1 mM cMCU- $\Delta$ NTD in 30  $\mu$ l reaction buffer were mixed with 0, 5, 7, 10 and 15 mM DTSSP, respectively, at room temperature. After 1 h, the reactions were quenched by adding 2  $\mu$ l of the stop solution (1 M Tris, pH 7.5). The reaction mixtures were analysed using SDS–PAGE, and visualized using silver stain.

**ITC analysis of Ru360 binding.** ITC was used to investigate whether cMCU- $\Delta$ NTD can interact with the known inhibitor Ru360 under the NMR sample condition. The protein samples used for ITC consisted of 20 mM MES pH 6.4, 75 mM sodium chloride, 4 mM foscholine-14 and 76  $\mu$ M cMCU- $\Delta$ NTD (monomer). The inhibitor solution contained 200  $\mu$ M Ru360 in 20 mM MES, pH 6.4, 75 mM sodium chloride and 4 mM foscholine-14. The titration protocol involved injecting 0.5  $\mu$ l inhibitor solution to 202  $\mu$ l protein solution for the first point and 1  $\mu$ l for each of the following 34 points (Extended Data Fig. 1c). Experiments were performed at 25 °C using Microcal ITC200 (GE Healthcare). The same ITC experiment was also performed for the cMCU- $\Delta$ NTD mutant with the S238A mutation.

**Negative stain EM analysis.** For EM analysis, the protein sample obtained from the size-exclusion chromatography above was subjected to two more rounds of size exclusion to achieve even higher homogeneity. The middle 20% of the elution peak was collected and subjected to another round of size-exclusion purification

under the same condition. The middle 20% of the elution from the second size exclusion run was collected and subjected to a third round of size exclusion. Finally, the middle 20% of the third elution was collected as the final sample for EM study.

To prepare sample for collecting negative stain EM images, 5  $\mu$ l of the cMCU- $\Delta$ NTD sample from the third size exclusion run (10  $\mu$ g ml<sup>-1</sup> protein in 20 mM MES, pH 6.5, 75 mM NaCl, 0.48 mM foscholine-14, 1 $\times$  protease inhibitor cocktail (Thermo) and 0.1 mM Na<sub>3</sub>N) was applied to a glow-discharged 400 mesh continuous carbon grid (Beijing Zhongjingkeyi Technology). The sample was then negatively stained with 0.75% (wt/vol) uranyl formate solution, and spontaneously dried at room temperature. The data were recorded on a Tecnai G2 Spirit BioTWIN transmission electron microscope (FEI) operated at 120 kV and equipped with an Eagle 4K  $\times$  4K CCD camera. We recorded 514 images at 110,000 microscope magnification with a pixel size of 1.05 Å per pixel. The defocus ranged from 1.0 to 1.5  $\mu$ m. For initial model building, Random Conical Tilt (RCT) image pairs were manually taken at tilt angles of 45° and 0°, respectively.

We boxed out 12,860 particles by using the e2boxer.py program in EMAN2.1 (ref. 29) (Extended Data Fig. 4a). Contrast transfer function parameters were determined for particles boxed out from each CCD image using EMAN1.9 (ref. 30) procedure ctf, followed by phase flipping using the applyctf program. The data were then low-pass filtered to 10 Å to enhance the image contrast for three-dimensional (3D) reconstruction<sup>31</sup>. Reference-free 2D analysis used the EMAN1.9 program refine2d.py, which revealed the existence of five-fold symmetry in the cMCU- $\Delta$ NTD sample (Extended Data Fig. 4b). The initial model was generated from RCT data using e2rct.py program in EMAN2.1. This model was further refined with the 12,860 untilted particles by using the refine program and calling the FRM2D image alignment kernel<sup>32,33</sup> in EMAN1.9. Initially, no symmetry was imposed in the 3D reconstruction process (Extended Data Fig. 4c), and subsequently the five-fold symmetry revealed by reference-free 2D analysis was imposed in the 3D reconstruction process. The final resolution was estimated at 18 Å by the 0.5 FSC criteria using the eotest program in EMAN1.9 (Extended Data Fig. 4d).

**Assignment of NMR resonances.** All NMR experiments were conducted at either 23 °C or 33 °C on Bruker spectrometers equipped with cryogenic probes. NMR spectra were processed using NMRPipe<sup>34</sup> and analysed using ccpNMR<sup>35</sup> and Xeasy<sup>36</sup>. Sequence-specific assignment of backbone <sup>1</sup>H<sup>N</sup>, <sup>15</sup>N, <sup>13</sup>C <sup>$\alpha$</sup> , <sup>13</sup>C <sup>$\beta$</sup>  and <sup>13</sup>C' chemical shifts was achieved using the TROSY versions of standard triple resonance experiments including HNCA, HN(CO)CA, HNCACB, HN(CA)CO and HNCO<sup>37,38</sup>. In addition, a 3D HSQC-NOESY-TROSY experiment with <sup>15</sup>N, <sup>15</sup>N and <sup>1</sup>H<sup>N</sup> evolution in the *t*<sub>1</sub>, *t*<sub>2</sub> and *t*<sub>3</sub> dimensions, respectively, was recorded with an NOE mixing time (*t*<sub>NOE</sub>) of 200 ms. These experiments were performed using multiple (<sup>15</sup>N, <sup>13</sup>C, <sup>2</sup>H)-labelled protein samples on a 600 MHz spectrometer at 33 °C. Multiple samples were used due to the poor stability of the protein at temperature >23 °C, that is, the sample began to show precipitation after ~7 days at 33 °C. Despite the problem, the higher temperature was used to obtain more favourable *T*<sub>2</sub> for triple resonance experiments. By combining the triple resonance data with the use of NOESY, we were able to confidently assign 91% of non-proline residues, although only 77% could be assigned if using only the triple resonance spectra. Protein aliphatic and aromatic resonances were assigned using a combination of 2D <sup>13</sup>C-edited HSQC, 3D <sup>15</sup>N-edited NOESY-TROSY (*t*<sub>NOE</sub> = 100 ms) and <sup>13</sup>C-edited NOESY-HSQC (*t*<sub>NOE</sub> = 150 ms) recorded on a 900 MHz spectrometer at both 23 °C and 33 °C. These experiments were performed using multiple (<sup>15</sup>N, <sup>13</sup>C)-labelled protein samples in which foscholine-14 was deuterated (Anatrace). The data sets recorded at two different temperatures provided complementary information. On average, the spectra at 23 °C show stronger peaks for the extramembrane regions but very weak peaks for the TM pore domain, especially the selectivity segment, possibly due to higher rigidity of the TM domain. At 33 °C, however, the TM resonances that were too weak to analyse in the 23 °C spectra became sufficiently intense.

**Assignment of NOEs.** Short-range NOEs used for defining local and secondary structures were assigned using the <sup>15</sup>N-edited NOESY-TROSY and <sup>13</sup>C-edited NOESY-HSQC recorded at two different temperatures as described above. Intermonomer NOEs between the backbone amide protons of a monomer and the non-exchangeable aliphatic or aromatic protons of its neighbouring monomers were assigned using a sample that was reconstituted with approximately 1:1 mixture of (<sup>15</sup>N, <sup>2</sup>H)-labelled cMCU- $\Delta$ NTD and (15% <sup>13</sup>C)-labelled cMCU- $\Delta$ NTD. The non-deuterated subunit was (15% <sup>13</sup>C)-labelled for recording the <sup>1</sup>H–<sup>13</sup>C HSQC spectrum (optimized for methyl groups) as internal aliphatic proton chemical shift reference while providing stereospecific assignment of leucine and valine methyl groups<sup>39</sup>. Mixing was done at the cell level before cell lysis, that is, the amounts of differently labelled cells were adjusted based on the protein expression level to ensure approximately 1:1 ratio of (<sup>15</sup>N, <sup>2</sup>H)- and (15% <sup>13</sup>C)-labelled cMCU- $\Delta$ NTD. In this sample, foscholine-14 was also deuterated. Recording a <sup>15</sup>N-edited

NOESY-TROSY using this sample allowed exclusive detection of NOE cross peaks between the  $^{15}\text{N}$ -attached protons of one monomer and aliphatic protons of adjacent monomers. The 3D  $^{15}\text{N}$ -edited NOESY-TROSY ( $\tau_{\text{NOE}} = 300\text{ ms}$ ) was recorded using this type of sample at both 23 °C and 33 °C. The intermonomer NOEs between the neighbouring CCHs and TMH2s (see main text) effectively defined the core of the cMCU- $\Delta\text{NTD}$  complex. The initial structural solution of the core assembly then allowed us to assign iteratively additional intra- and intermonomer long-range NOEs between the aliphatic and aromatic protons in  $^{15}\text{N}$ -edited NOESY-TROSY and  $^{13}\text{C}$ -edited NOESY-HSQC recorded at both 23 °C and 33 °C. These long-range NOEs subsequently defined packing of TMH1 against TMH2, packing of H1 against CCH, as well as interaction between IJMH and OJMH.

**Structure calculation.** Structures were calculated using the program XPLOR-NIH<sup>40</sup>. The local and secondary structures of the monomer were first defined using short-range NOE restraints and backbone dihedral restraints derived from chemical shifts (using TALOS+<sup>41</sup>). A total of 10 monomer structures were calculated using a standard simulated annealing protocol. Five copies of the lowest-energy monomer structure were used to construct an initial model of the pentamer using the intermonomer NOE restraints assigned for the pore-forming TMH2 and CCH helices. For each intermonomer restraint between two adjacent monomers, five identical distance restraints were assigned respectively to all pairs of neighbouring monomers to satisfy the condition of C5 rotational symmetry (as indicated by the EM reconstruction). The assembled pentamer was used as a starting model to guide assignment of more long-range NOEs in the  $^{15}\text{N}$ -edited and  $^{13}\text{C}$ -edited NOE spectra recorded at 23 °C and 33 °C. This process was repeated iteratively until the calculated models converged to backbone r.m.s.d. of  $<1\text{ \AA}$  (not including the disordered regions). In each of the iterations, the starting model was refined against dihedral restraints and local and long-range intra- and intermonomer NOE restraints using a simulated annealing protocol in which the bath was cooled from 1,000 to 100 K. The NOE restraints were enforced by flat-well harmonic potentials, with the force constant ramped from 25 to 50 kcal mol<sup>-1</sup> Å<sup>-2</sup> during annealing. For the defined helical regions, backbone dihedral angle restraints ( $\phi = -60^\circ$ ,  $\psi = -40^\circ$ ) were applied, all with a flat-well ( $\pm 10^\circ$ ) harmonic potential with force constant ramped from 15 to 30 kcal mol<sup>-1</sup> rad<sup>-2</sup>. In the final round of refinement, a total of 150 structures were calculated and 15 low-energy structures were selected as the structural ensemble. Ramachandran plot statistics for the structure ensemble, calculated using PROCHECK<sup>42</sup>, are as follows: most favoured (86.3%), additionally allowed (11.0%), generously allowed (1.8%) and disallowed (1.0%). Restraint and refinement statistics are shown in Extended Data Table 1.

#### Characterization of the oligomeric state of the CCH peptide by crosslinking.

To examine whether the C-terminal CCH domain is able to oligomerize by itself in water, we investigated the oligomeric state of a CCH-containing peptide corresponding to residues 288–316 of cMCU (plus the C-terminal L and E as in the cMCU- $\Delta\text{NTD}$  construct) by chemical crosslinking using DTSSP (3,3'-dithiobis (sulfosuccinimidyl propionate) (Thermo Scientific). Approximately 0.4 mg synthesized peptide (Pepmic Co., Ltd) was dissolved in 1 ml of reaction buffer (25 mM sodium phosphate, pH 7.5, and 50 mM NaCl) to a concentration of 0.1 mM. A stock solution of 30 mM DTSSP in water was prepared for addition to the peptide solution. Four samples, each containing 0.1 mM peptide in 30  $\mu\text{l}$  reaction buffer were incubated with 0, 1, 3 and 5 mM DTSSP at room temperature. After 1 h, the reactions were quenched by adding 1  $\mu\text{l}$  of stop solution (1 M Tris, pH 7.5). The reaction mixtures were analysed using SDS-PAGE. The SDS-PAGE results show that at DTSSP concentrations of 3 and 5 mM, the peptides could be crosslinked up to pentamers (Extended Data Fig. 7), indicating that the synthesized peptide can form pentamers in water.

**Functional mutagenesis of HsMCU.** *Material.* Reagents were obtained from the following sources: anti-Flag M2 affinity gel from Sigma (A2220), ATP5A antibody

from Abcam (ab14748), Oregon Green 488 BAPTA-6 F from Life Technologies (O23990).

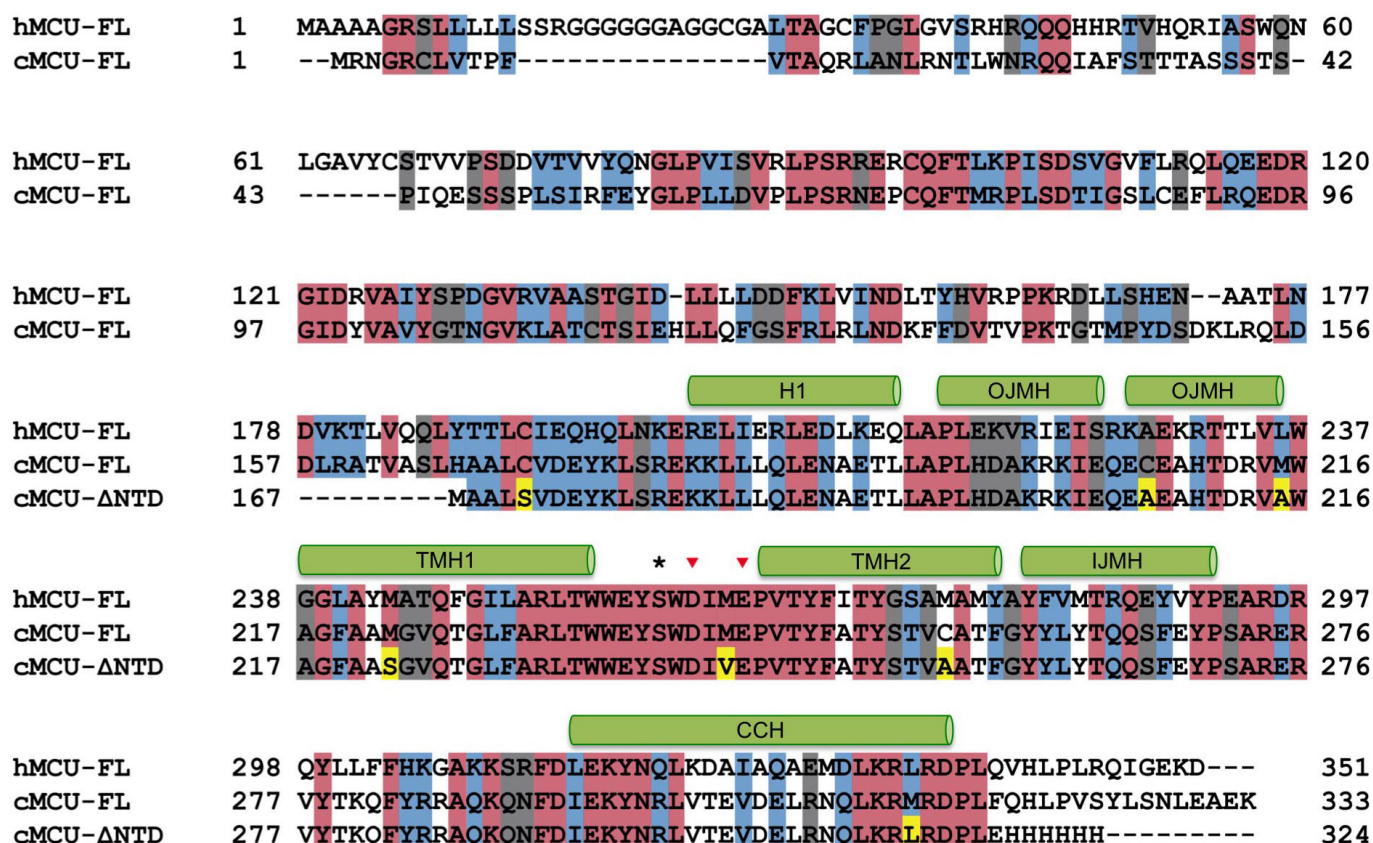
**Cell culture and cell lines.** MCU knockout cell lines were generated as previously described<sup>7</sup>. Cells were infected with lentivirus for stable expression of Flag-tagged proteins and selected with 100  $\mu\text{g ml}^{-1}$  puromycin. Cell lines have been tested for mycoplasma contamination on a monthly basis using MycoAlert Mycoplasma Detection Kit (Lonza, LT07-418) and are free of mycoplasma contamination. The parental HEK-293T cell line has the following STR profile: TH01 (7, 9, 3); D21S11 (28, 29, 30, 2); D5S818 (7, 8, 9); D13S317 (11, 12, 13, 14, 15); D7S820 (11); D16S539 (9, 13); CSF1PO (11, 12, 13); Amelogenin (X); vWA (16, 18, 19, 20); TPOX (11). This profile matches 100% to HEK-293T cell line profile (ATCC, CRL-3216) if the Alternative Master's algorithm is used, and 83% if the Tanabe algorithm is used. MCU knockout cell line was generated from this parental cell line by single cell cloning after MCU was knocked out using TALEN technology.

**Calcium uptake in permeabilized HEK-293T cells.** Calcium uptake assay in permeabilized HEK-293T cells was done as previously described<sup>7</sup>. Rate of calcium uptake is defined as the slope of the linear portion of the calcium uptake curve (between 20 and 30 s). In each experiment, calcium uptake rates of samples were normalized to the calcium uptake rate of a randomly selected wild-type HEK-293T cell uptake rate.

**HsMCU- $\Delta\text{NTD}$  cloning and expression.** For expression and mitochondrial targeting of HsMCU- $\Delta\text{NTD}$ , cDNA that corresponds to HsMCU amino acids 187–351 was fused to HsMCU mitochondrial targeting signal (amino acids 1–56) and cloned into pLYS1 vector as described previously<sup>7</sup>.

28. Slotboom, D. J., Duurkens, R. H., Olieman, K. & Erkens, G. B. Static light scattering to characterize membrane proteins in detergent solution. *Methods* **46**, 73–82 (2008).
29. Tang, G. *et al.* EMAN2: an extensible image processing suite for electron microscopy. *J. Struct. Biol.* **157**, 38–46 (2007).
30. Ludtke, S. J., Baldwin, P. R. & Chiu, W. EMAN: semiautomated software for high-resolution single-particle reconstructions. *J. Struct. Biol.* **128**, 82–97 (1999).
31. Guo, X. *et al.* Structural insight into autoinhibition and histone H3-induced activation of DNMT3A. *Nature* **517**, 640–644 (2015).
32. Cong, Y., Kovacs, J. A. & Wriggers, W. 2D fast rotational matching for image processing of biophysical data. *J. Struct. Biol.* **144**, 51–60 (2003).
33. Cong, Y. *et al.* Fast rotational matching of single-particle images. *J. Struct. Biol.* **152**, 104–112 (2005).
34. Delaglio, F. *et al.* NMRPipe: a multidimensional spectral processing system based on UNIX pipes. *J. Biomol. NMR* **6**, 277–293 (1995).
35. Vranken, W. F. *et al.* The CCPN data model for NMR spectroscopy: development of a software pipeline. *Proteins* **59**, 687–696 (2005).
36. Bartels, C., Xia, T. H., Billeter, M., Guntert, P. & Wuthrich, K. The program XEASY for computer-supported NMR spectral analysis of biological macromolecules. *J. Biomol. NMR* **6**, 1–10 (1995).
37. Kay, L. E., Ikura, M., Tschudin, R. & Bax, A. Three-dimensional triple resonance NMR spectroscopy of isotopically enriched proteins. *J. Magn. Reson.* **213**, 423–441 (1990).
38. Salzmann, M., Wider, G., Pervushin, K. & Wuthrich, K. Improved sensitivity and coherence selection for [ $^{15}\text{N}$ ,  $^1\text{H}$ ]-TROSY elements in triple resonance experiments. *J. Biomol. NMR* **15**, 181–184 (1999).
39. Szyperski, T., Neri, D., Leiting, B., Otting, G. & Wuthrich, K. Support of  $^1\text{H}$  NMR assignments in proteins by biosynthetically directed fractional  $^{13}\text{C}$ -labeling. *J. Biomol. NMR* **2**, 323–334 (1992).
40. Schwieters, C. D., Kuszewski, J., Tjandra, N. & Clore, G. M. The Xplor-NIH NMR molecular structure determination package. *J. Magn. Reson.* **160**, 65–71 (2003).
41. Shen, Y., Delaglio, F., Cornilescu, G. & Bax, A. TALOS+: a hybrid method for predicting protein backbone torsion angles from NMR chemical shifts. *J. Biomol. NMR* **44**, 213–223 (2009).
42. Laskowski, R. A., MacArthur, M. W., Moss, D. S. & Thornton, J. W. PROCHECK: a program to check the stereochemical quality of protein structures. *J. Appl. Crystallogr.* **26**, 283–291 (1993).

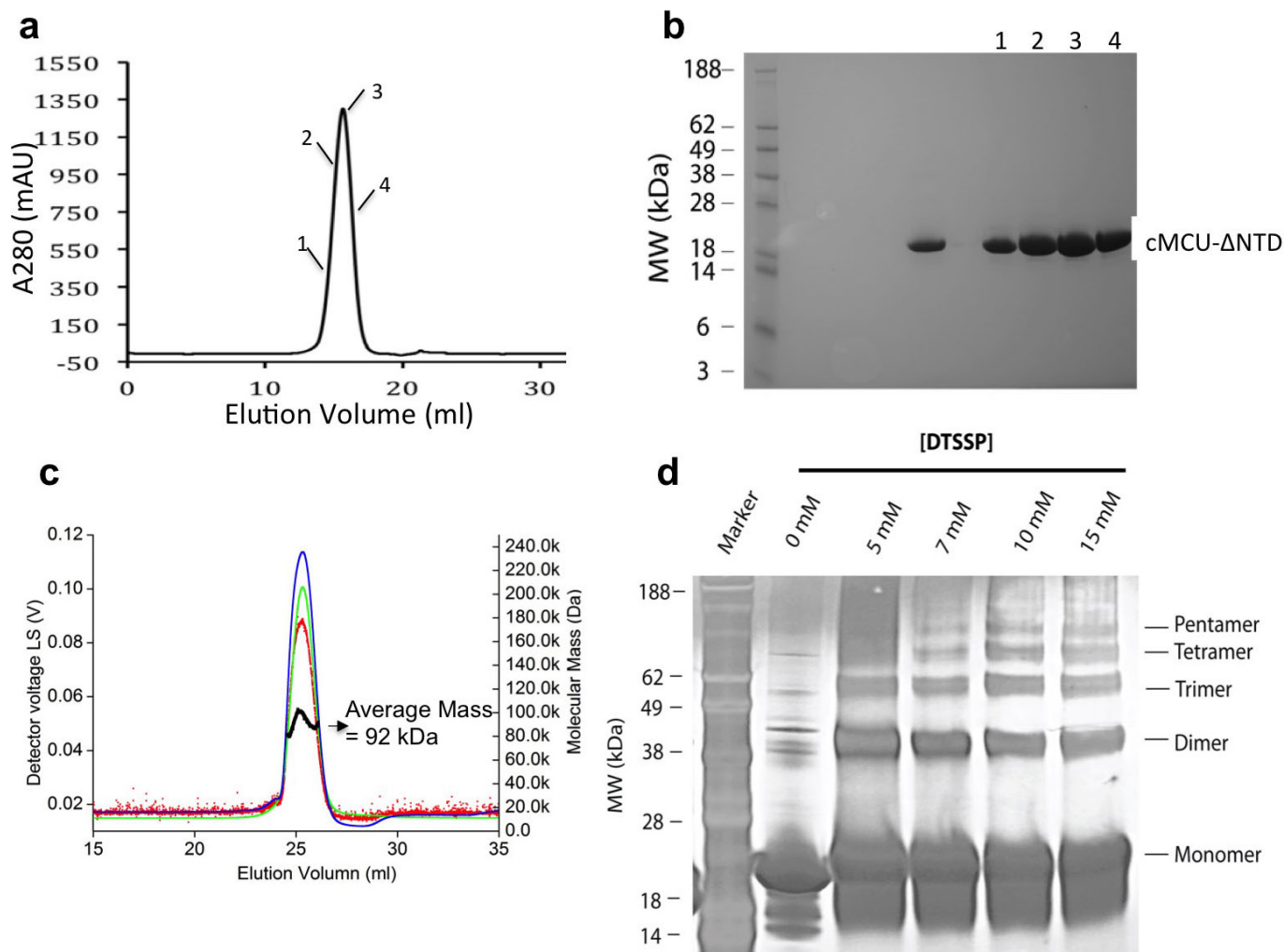




Extended Data Figure 1 | Multiple sequence alignment of the full-length hMCU, full-length cMCU and cMCU-ΔNTD. Residues that are invariant in all three sequences are shaded in red. Partially conserved and much less conserved residues are shaded in blue and grey, respectively. The mutations introduced in cMCU-ΔNTD are shaded in yellow. D and E

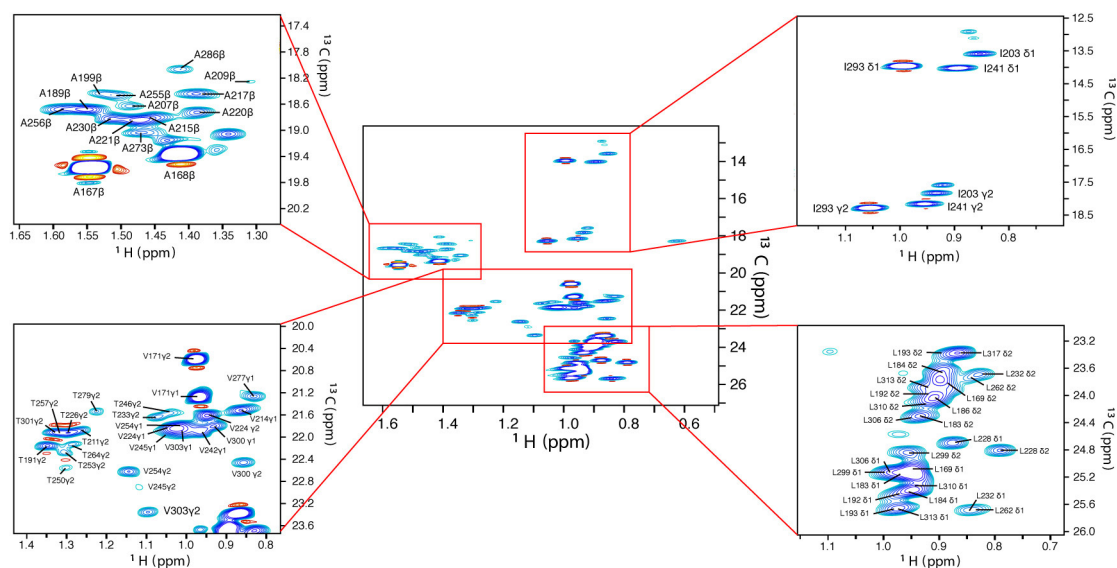
in the DXXE motif are indicated by downward arrow. The serine residue shown to be involved in Ru360 inhibition is indicated by an asterisk<sup>4</sup>. Helical segments, as determined by NMR in this study, are indicated by cylinders and labelled as in the main text. The accession numbers for hMCU and cMCU are NM\_138357.1 and NP\_500892.1, respectively.



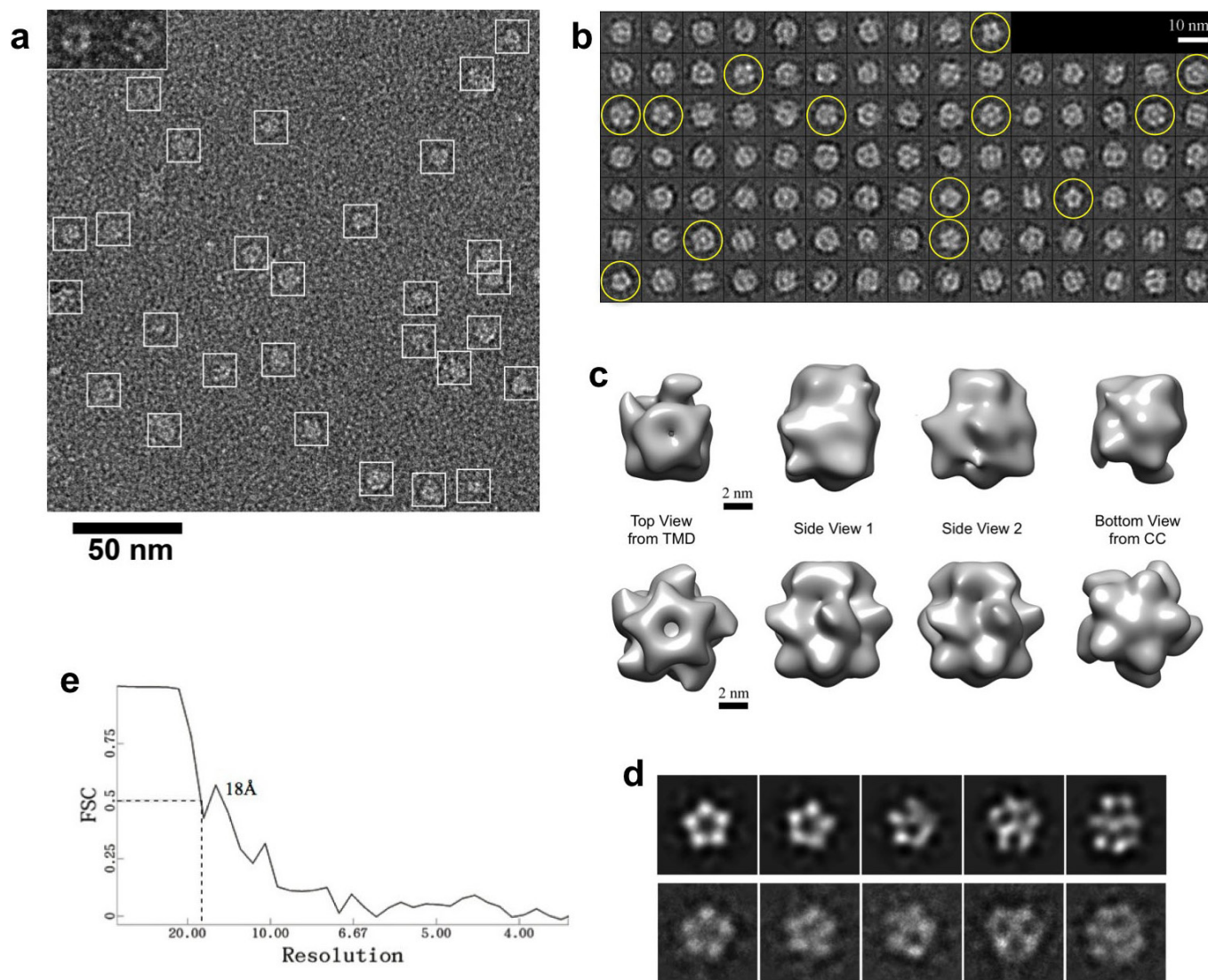


**Extended Data Figure 2 | Biochemical analysis of cMCU-ΔNTD oligomeric state.** **a**, Elution peak of cMCU-ΔNTD from Superdex 200 10/300 GL column in 20 mM MES, pH 6.4, 75 mM NaCl, 0.48 mM foscholine-14, 0.3 mM NaN<sub>3</sub> and 2 mM EDTA. **b**, SDS-PAGE analysis of the elution peak showing sample purity >95%. **c**, SEC-MALS analysis of the diluted NMR sample of cMCU-ΔNTD. The SEC-MALS/UV/refractive index measurement was used to determine cMCU-ΔNTD molecular mass based on the three-detector method<sup>28</sup>. In this method, since the foscholine-14 detergent does not have UV absorption at 280 nm, the protein mass is directly calculated without correcting for the bound micelle. Chromatograms show the readings from the light scattering at 90° (red), refractive index (blue), and UV (green) detectors. The left and right axes represent the light scattering detector reading and molecular mass, respectively. The black curve represents the calculated molecular mass,

and the average mass of the elution peak of cMCU-ΔNTD is 92 kDa. Note the ~10 ml difference in elution volumes between **a** and **c** is due to the volume of solution feeding from SEC to MALS. **d**, SDS-PAGE analysis of chemical crosslinking of the diluted cMCU-ΔNTD NMR sample. The reaction mixture contains 0.1 mM cMCU-ΔNTD (monomer), 3 mM foscholine-14, and various amounts of DTSSP. The reactions were quenched after 1 h by the addition of 2 μl of 1 M Tris, pH 7.5. The quenched samples were loaded to 12% Bis-Tris gel (Novex Life Technologies). The gel was silver stained using the standard protocol. The five lanes to the right of the molecular mass marker correspond to DTSSP concentrations of 0, 5, 7, 10 and 15 mM. The band that corresponds to a pentamer showed the most obvious increase in intensity as a function of [DTSSP].

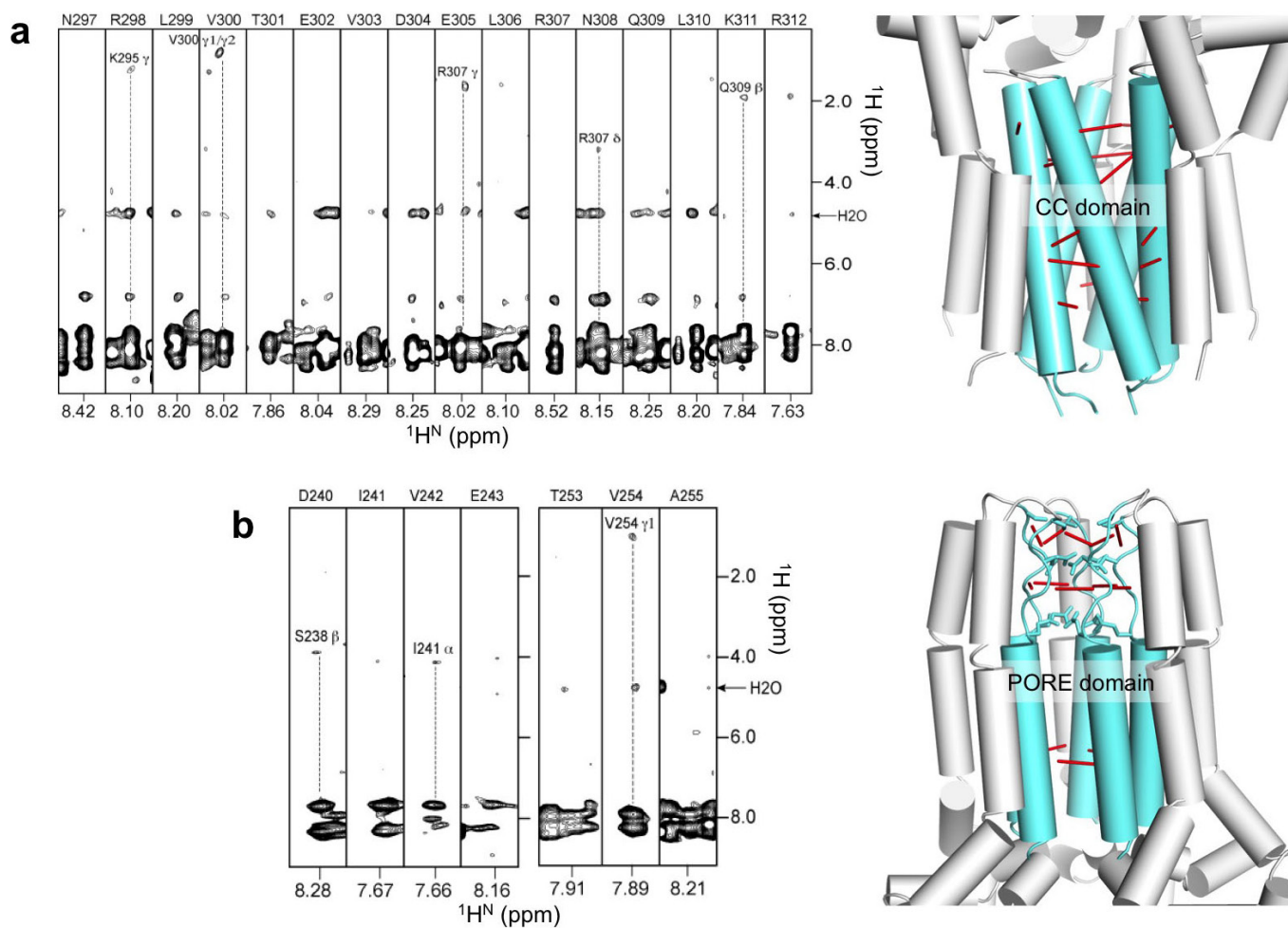


**Extended Data Figure 3 | The eMCU- $\Delta$ NTD methyl group resonances with residue specific assignment.** The  $^1\text{H}$ - $^{13}\text{C}$  HSQC was recorded with 28 ms constant-time  $^{13}\text{C}$  evolution at 900 MHz.



**Extended Data Figure 4 | Single particle EM of the cMCU- $\Delta$ NTD oligomeric complex.** **a**, Typical image of the cMCU- $\Delta$ NTD oligomers negatively stained with uranyl formate. The bar corresponds to 50 nm in length. A selected subset of cMCU- $\Delta$ NTD particles is highlighted with white circles. Shown in the top left corner are a typical top/bottom view and a typical side view of the selected particles. **b**, Gallery of 100 out of 202 reference-free 2D class averages of the particles, which revealed the existence of five-fold symmetry in the complex (pentagon shapes marked by yellow circles). **c**, Comparison of different views of the 3D EM

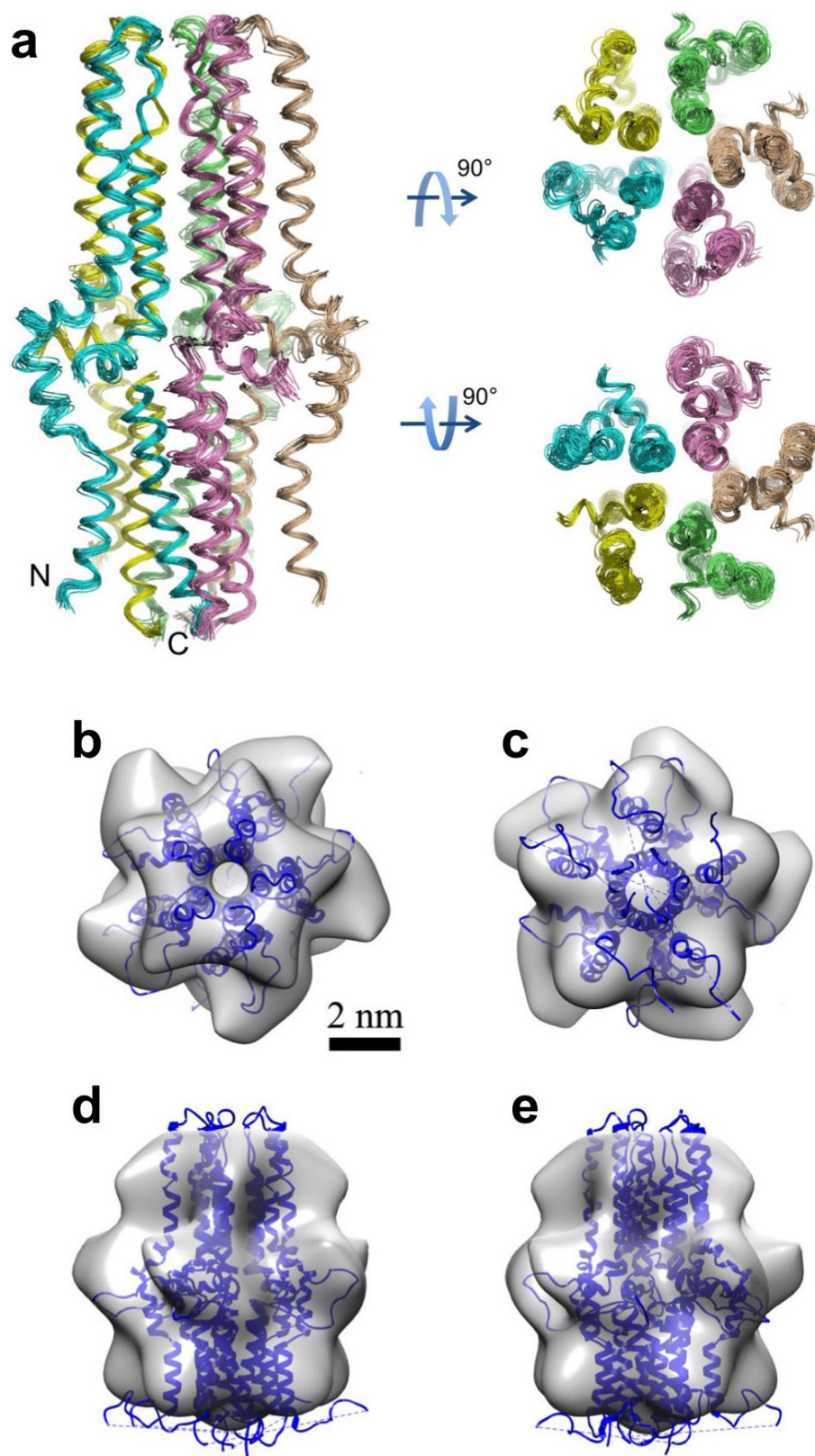
density map reconstructed without enforcing C5 symmetry (top) with the corresponding views of the map reconstituted with symmetry (bottom). The largest discrepancies are in the middle bulge region between the TM and CC domains, possibly due to the lack of rigid structure in the large L2 loop. **d**, Comparison of the 2D projections (top) from the 3D EM density map with the corresponding reference-free 2D classes. **e**, Estimation of resolution of the final 3D reconstruction. The Fourier shell correlation (FSC) suggests a resolution of  $\sim 18$  Å using the 0.5 criterion.



**Extended Data Figure 5 | Intermonomer NOEs from mixed isotope labelled sample.** Examples are taken from the 3D  $^{15}\text{N}$ -edited NOESY-TROSY of the mixed labelled sample containing 1:1 mixture of ( $^{15}\text{N}$ ,  $^2\text{H}$ )-labelled cMCU- $\Delta\text{NTD}$  and (15%  $^{13}\text{C}$ )-labelled cMCU- $\Delta\text{NTD}$ . **a**, Sample  $^1\text{H}$ - $^1\text{H}$  strips at various  $^{15}\text{N}$  chemical shifts showing intermonomer NOEs between backbone amide proton and aliphatic protons for the C-terminal

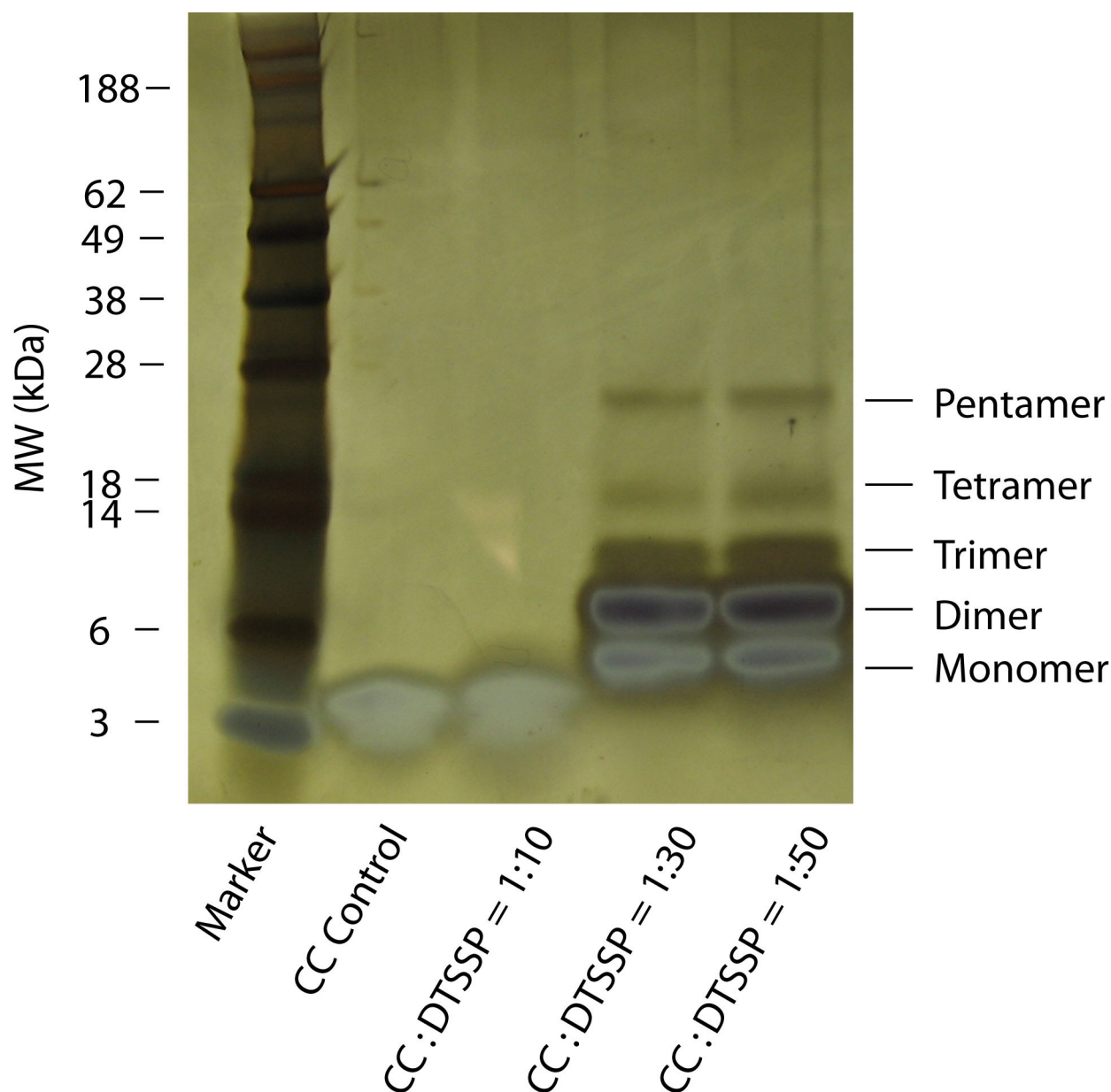
CCH domain. The NOE spectrum was recorded at 23 °C at 900 MHz. **b**, Sample strips showing intermonomer NOEs within the TMH2 pore as well as the selectivity filter region. The NOE spectrum was recorded at 33 °C at 900 MHz. On the right of each panel, intermonomer NOEs in the context of the structure are shown as red lines.





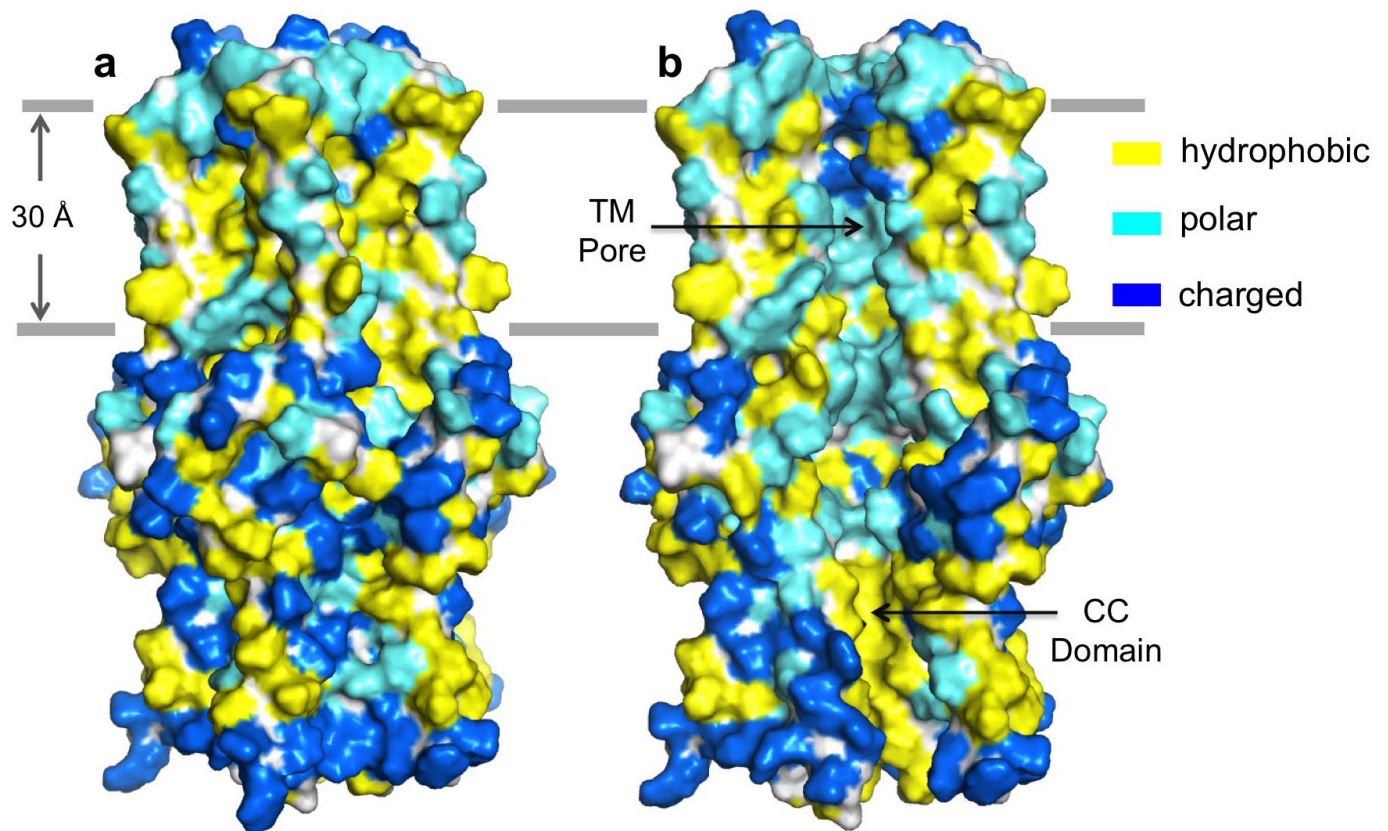
**Extended Data Figure 6 | Structural ensemble of the cMCU- $\Delta$ NTD pentamer derived from NMR restraints and fitting to the EM density map.** **a**, Ensemble of the 15 lowest-energy structures calculated using NMR-derived structural restraints (see Extended Data Table 1). The unstructured loops L1 (residues 166–179) and L2 (residues 272–292) are not shown for clarity. **b**, The NMR structure of cMCU- $\Delta$ NTD without the loop regions L1 and L2 was fitted to the EM volume using rigid body

fitting (the ‘fit’ tool in Chimera). The L1 and L2 are however included for display. Top view from the intermembrane space side. **c**, Bottom view from the matrix side. **d**, **e**, Two different side views. Note that the loop regions appear disordered due to the lack of NMR-derived structural restraints. This does not necessarily mean that they do not acquire any stable conformation. The presumed detergent molecules around the membrane-embedded region are not taken into account in this fit.



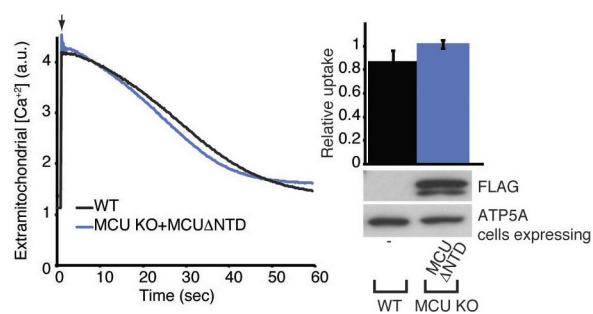
**Extended Data Figure 7 | SDS-PAGE analysis of chemical crosslinking of the CCH peptide.** The reaction mixtures containing 0.1 mM peptide (cMCU residues 288–316 plus the C-terminal L and E as in the cMCU- $\Delta$ NTD construct) and various amounts of DTSSP were quenched after 1 h of reaction by the addition of 1  $\mu$ l of 1 M Tris, pH 7.5. The quenched

samples were loaded to 12% Bis-Tris gel (Novex Life Technologies). The gel was silver stained using the standard protocol. The four lanes to the right of the molecular mass marker correspond to DTSSP:CCH ratios of 0, 10:1, 30:1 and 50:1.



**Extended Data Figure 8 | Surface representation for revealing the surface-exposed and core amino acid properties of the cMCU- $\Delta$ NTD pentamer.** Hydrophobic, polar and charged residues are shown in yellow, cyan and blue, respectively. The hydrophobic residues include A, I, L, F, V, P and G, the polar residues include Q, N, H, S, T, Y, C, M and W, and

the charged residues include K, R, D and E. The solid lines indicate the hydrophobic core boundaries of the presumed lipid bilayer. **a**, Pentamer with unstructured loops removed. **b**, The same view as in **a** but with the front subunit removed to reveal the core.



**Extended Data Figure 9 | Deletion of the NTD (amino acids 58–186) in HsMCU (HsMCU- $\Delta$ NTD) does not impair its function.** Representative traces of  $Ca^{2+}$  uptake in digitonin-permeabilized cells after addition of  $50\mu M$   $CaCl_2$  are shown on the left. The bar graph shows the rate of  $Ca^{2+}$  uptake relative to wild-type HEK-293T cells (mean  $\pm$  s.d.,  $n = 4$ ). Cell lysates were analysed by immunoblotting using an anti-Flag antibody to detect expression of MCU protein. ATP5A was used as loading control.



Extended Data Table 1 | NMR and refinement statistics for protein structures

cMCU-ΔNTD	
<b>NMR distance and dihedral constraints</b>	
Distance constraints	
Total NOE	2440
Intra-residue	130
Inter-residue	2310
Sequential ( $ i-j  = 1$ )	1280
Medium-range ( $ i-j  < 4$ )	660
Long-range ( $ i-j  > 5$ )	150
Intermolecular	220
Hydrogen bonds	0
Total dihedral angle restraints	990
phi	495
psi	495
<b>Structure statistics</b>	
Violations (mean and s.d.)	
Distance constraints (Å)	$0.231 \pm 0.002$
Dihedral angle constraints (°)	$1.485 \pm 0.034$
Max. dihedral angle violation (°)	9.511
Max. distance constraint violation (Å)	1.487
Deviations from idealized geometry	
Bond lengths (Å)	$0.034 \pm 0.000$
Bond angles (°)	$1.777 \pm 0.006$
Impropers (°)	$1.150 \pm 0.016$
Average pairwise r.m.s.d.** (Å)	
Heavy	1.503
Backbone	0.920

The numbers of constraints are summed over all five subunits. Backbone  $\phi$  and  $\psi$  restraints of  $-60^\circ$  and  $-40^\circ$ , respectively, were assigned for regions of the protein confirmed to be helical based on local NOEs and TALOS<sup>+</sup><sup>39</sup>. Statistics are calculated and averaged over an ensemble of the 15 lowest-energy structures out of 150 calculated structures. The precision of the atomic coordinates is defined as the average r.m.s.d. between the 15 final structures and their mean coordinates. The calculation includes only the structured regions of the protein: residues 180–192, 194–271 and 292–316.

# Extra-helical binding site of a glucagon receptor antagonist

Ali Jazayeri<sup>1\*</sup>, Andrew S. Doré<sup>1\*</sup>, Daniel Lamb<sup>1\*</sup>, Harini Krishnamurthy<sup>1\*</sup>, Stacey M. Southall<sup>1</sup>, Asma H. Baig<sup>1</sup>, Andrea Bortolato<sup>1</sup>, Markus Koglin<sup>1</sup>, Nathan J. Robertson<sup>1</sup>, James C. Errey<sup>1</sup>, Stephen P. Andrews<sup>1</sup>, Iryna Teobald<sup>1</sup>, Alastair J. H. Brown<sup>1</sup>, Robert M. Cooke<sup>1</sup>, Malcolm Weir<sup>1</sup> & Fiona H. Marshall<sup>1</sup>

Glucagon is a 29-amino-acid peptide released from the  $\alpha$ -cells of the islet of Langerhans, which has a key role in glucose homeostasis<sup>1</sup>. Glucagon action is transduced by the class B G-protein-coupled glucagon receptor (GCGR), which is located on liver, kidney, intestinal smooth muscle, brain, adipose tissue, heart and pancreas cells, and this receptor has been considered an important drug target in the treatment of diabetes. Administration of recently identified small-molecule GCGR antagonists in patients with type 2 diabetes results in a substantial reduction of fasting and postprandial glucose concentrations<sup>2</sup>. Although an X-ray structure of the transmembrane domain of the GCGR<sup>3</sup> has previously been solved, the ligand (NNC0640) was not resolved. Here we report the 2.5 Å structure of human GCGR in complex with the antagonist MK-0893 (ref. 4), which is found to bind to an allosteric site outside the seven transmembrane (7TM) helical bundle in a position between TM6 and TM7 extending into the lipid bilayer. Mutagenesis of key residues identified in the X-ray structure confirms their role in the binding of MK-0893 to the receptor. The unexpected position of the binding site for MK-0893, which is structurally similar to other GCGR antagonists, suggests that glucagon activation of the receptor is prevented by restriction of the outward helical movement of TM6 required for G-protein coupling. Structural knowledge of class B receptors is limited, with only one other ligand-binding site defined—for the corticotropin-releasing hormone receptor 1 (CRF<sub>1</sub>R)—which was located deep within the 7TM bundle<sup>5</sup>. We describe a completely novel allosteric binding site for class B receptors, providing an opportunity for structure-based drug design for this receptor class and furthering our understanding of the mechanisms of activation of these receptors.

To obtain a high-resolution structure of the human GCGR transmembrane domain (TMD), a thermostabilized receptor (StaR) was generated<sup>6–8</sup> containing 11 amino-acid substitutions. To facilitate crystallization further, the extracellular domain was removed from the N terminus (residues 2–135), and the C terminus was truncated by 60 residues (residues 418–477). Finally, T4-lysozyme (T4L) was inserted into intracellular loop (ICL)2 between Leu255 and Pro259 (Fig. 1a), resulting in the construct designated GCGR-StaR(136–417)–T4L. The structure was solved in the presence of the antagonist MK-0893 (Fig. 1b and Extended Data Table 1). The construct modifications did not alter the antagonist-binding properties of the receptor compared with wild type (Extended Data Table 2).

The core fold of the receptor features the canonical 7TM helices (TM1–TM7) (Fig. 1c) in a similar conformation to the previously published GCGR structure<sup>3</sup>. Continuous density is observed for intracellular loops ICL1 and ICL3, while extracellular loop (ECL)2 adopts a conformation capping the entrance to the orthosteric site. In contrast with the previously published GCGR structure<sup>3</sup>, the N terminus of TM5 unwinds by one helical turn, permitting ECL2 to stretch across to the

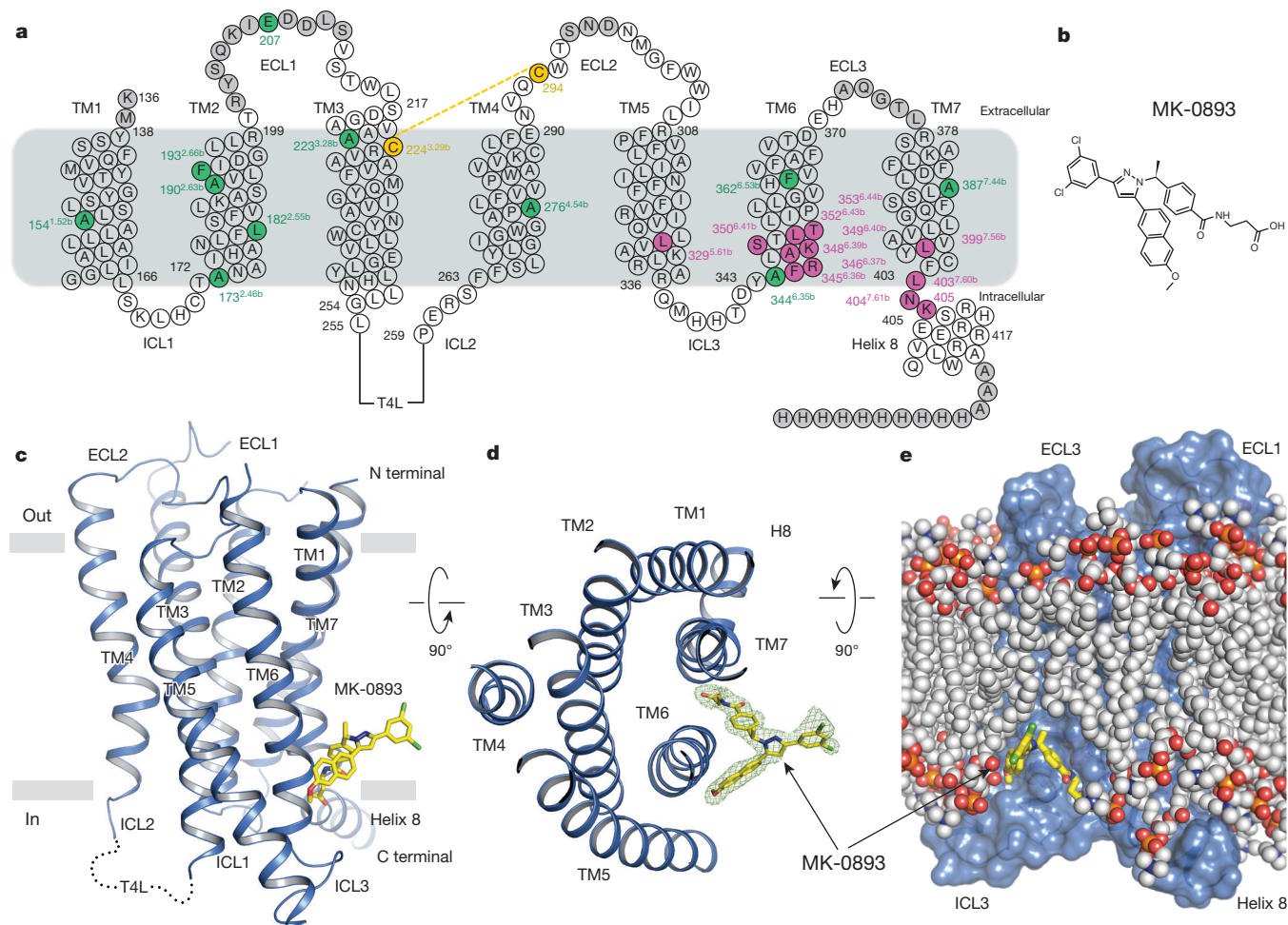
central axis of the TM helical bundle, mediating interactions from TM3 across to TM6 and TM7 while maintaining the conserved disulfide bond between Cys224<sup>3.29b</sup> and Cys294 (numbers in superscript refer to the modified Ballesteros numbering system for class B GPCRs<sup>9–11</sup>). The highly conserved sequence motif GWGxP in TM4 of class B receptors has an important structural role supplying interactions stabilizing the configuration of TM2, TM3 and TM4 (ref. 5). In this GCGR structure, TM4 bulges at Gly271<sup>4.49b</sup> and, along with Pro275<sup>4.53b</sup>, they disrupt intra-helical hydrogen bonding and result in positioning of Trp272<sup>4.50b</sup> towards TM2 and TM3, which forms a hydrogen bond with Asn179<sup>2.52b</sup> on TM2 in an analogous fashion to CRF<sub>1</sub>R. In addition, a hydrogen bond from the side chain of Tyr233<sup>3.38b</sup> to the backbone carbonyl of Trp272<sup>4.50b</sup> further strengthens the TM4–TM3 interaction.

Unexpectedly, strong and unambiguous density is observed for the MK-0893 antagonist outside the 7TM helical bundle (Fig. 1c–e), straddling TM6 from within the lipid bilayer. TM6 then acts to divide the binding site into two distinct regions: a hydrophobic interface with TM5, and a polar cleft towards TM7. The different physicochemical properties of this bipartite antagonist pocket correspond to the dual hydrophilic/hydrophobic nature of the ligand. The apolar methoxynaphthalene moiety of the small molecule makes hydrophobic contacts in the TM5–TM6 interface with Leu329<sup>5.61b</sup>, Phe345<sup>6.36b</sup>, Leu352<sup>6.43b</sup>, Thr353<sup>6.44b</sup> and the alkyl chain of Lys349<sup>6.40b</sup> (Fig. 2a). On the opposite site, within the TM6–TM7 cleft, MK-0893 participates in a network of polar contacts: the ligand amide group hydrogen bonds with Lys349<sup>6.40b</sup> and Ser350<sup>6.41b</sup>, while the carboxyl group forms a salt bridge with Arg346<sup>6.37b</sup>. This moiety also makes additional polar interactions with Asn404<sup>7.61b</sup>, the backbone of Lys405 (located between TM7 and H8) and a water-mediated hydrogen bond with Ser350<sup>6.41b</sup> and Leu399<sup>7.56b</sup> (Fig. 2a). The phenylethylpyrazole core of the molecule makes further interactions with TM6, in particular with Thr353<sup>6.44b</sup> and Lys349<sup>6.40b</sup>. The position of the pyrazole moiety parallel to the membrane provides the two ligand vectors towards the bipartite antagonist sub-pockets. A third vector starting from the ligand pyrazole ring leads to the dichlorophenyl group, which makes a crystal contact to TM4 of a symmetry mate (see Extended Data Fig. 1). Molecular dynamics simulations of MK-0893 binding to the wild-type receptor (outside the constraints of the crystal system) demonstrate that the receptor–ligand interaction is stable, and MK-0893 remains at the cytoplasmic membrane boundary with the carboxyl moiety able to interact with intracellular solvent and with interactions between the ligand and receptor, involving Arg346<sup>6.37b</sup>, Lys349<sup>6.40b</sup>, Asn404<sup>7.61b</sup> and Lys405, maintained (Extended Data Fig. 2). Sequence conservation analysis demonstrates that, with the exception of Thr353<sup>6.44b</sup> and Phe345<sup>6.36b</sup>, the other key binding-site residues show good conservation across other members of human class B receptors (Extended Data Fig. 5).

Tritium-labelled MK-0893 was prepared and used to characterize the ligand-binding site. Membrane fractions prepared from HEK293T

<sup>1</sup>Heptares Therapeutics Ltd, BioPark, Broadwater Road, Welwyn Garden City, Hertfordshire AL7 3AX, UK.

\*These authors contributed equally to this work.



**Figure 1 | Structure of GCGR and the MK-0893 allosteric binding site.** **a**, Crystallization construct showing stabilizing mutations (green), binding site residues (pink), disordered residues not located (grey), and the disulfide bond between Cys224<sup>3,29b</sup> and Cys294 (yellow line). **b**, Chemical structure of MK-0893. **c, d**, Ribbon representation of GCGR (blue), viewed parallel to the membrane (**c**) and from extracellular space (**d**). The position

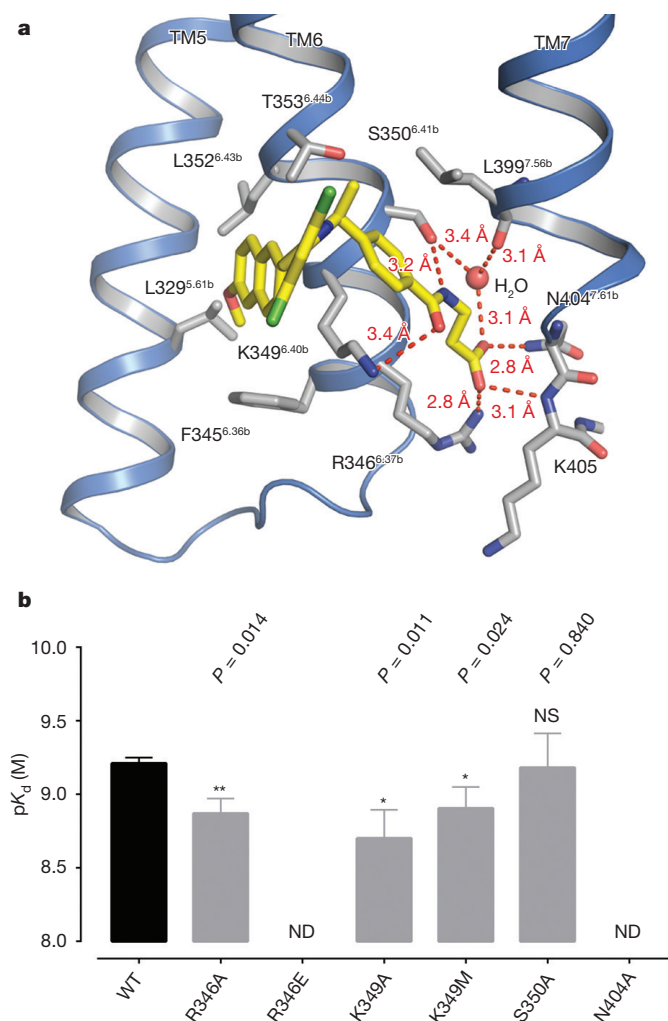
of the T4L insertion to ICL2 is indicated. MK-0893 in stick representation with carbon, nitrogen, oxygen and chlorine atoms coloured yellow, blue, red and green, respectively,  $F_o - F_c$  OMIT density contoured at 2.5 $\sigma$ . **e**, Surface representation of GCGR embedded within the membrane (20 ns molecular dynamics simulation), lipids in proximity of MK-0893 removed for clarity.

cells expressing the wild-type GCGR were used in saturation binding analysis. Saturation binding was monophasic and best fitted to a one-site model with a linear Scatchard plot, consistent with the presence of a single high-affinity binding site (Fig. 3a). Competition with [<sup>3</sup>H]MK-0893 was used to characterize the binding of a number of reported GCGR antagonists (see Extended Data Fig. 3 for details). These compounds were selected to represent molecules that exhibit chemical similarity to MK-0893 (for example, NNC0640, Cpd-01, 02, 03 and 04), as well as those that are chemically distinct (for example, Cpd-05 and 06). Consistently, NNC0640, Cpd-01, 02, 03 and 04 were able to fully compete with [<sup>3</sup>H]MK-0893 binding, indicating that these compounds share the same binding site (Fig. 3b–f). By contrast, Cpd-05 and Cpd-06 were not competitive with [<sup>3</sup>H]MK-0893, indicating that these compounds bind to a different site (Fig. 3g, h). Furthermore, [<sup>3</sup>H]-MK-0893 was not displaced by glucagon or the related peptide antagonist des-His1-[Glu9]-glucagon (Fig. 3i, j), which bind at the orthosteric site. Interestingly, NNC0640 was the ligand used in the crystallization of the first reported GCGR structure<sup>3</sup>; however, the position of the ligand was not resolved. A clear peak is observed in the electron density map of GCGR–NNC0640 between TM6 and TM7 towards the intracellular side of the receptor. Although this was modelled as a polyethylene glycol (PEG) molecule in the reported structure, superposition with the GCGR structure reported here demonstrates that the amide moiety and carboxyl function of MK-0893 directly overlays with this peak

(Extended Data Fig. 4). Given the chemical similarity of these ligands, coupled with our competition data, it is likely that this constitutes residual signal from NNC0640 binding in an analogous position to MK-0893 on GCGR rather than being a PEG molecule.

To confirm the allosteric pocket identified in the structure, single point mutations were made to residues in the binding site and the binding of [<sup>3</sup>H]MK-0893 was subsequently assessed. The mutations were introduced in the full-length wild-type receptor with a C-terminal enhanced green fluorescent protein (eGFP) tag. The presence of the eGFP tag has no impact on MK-0893 binding or glucagon activation of the receptor (data not shown). For this analysis, Arg346<sup>6,37b</sup> was mutated to alanine or glutamic acid. Lys349<sup>6,40b</sup> was mutated to either alanine or methionine. Ser350<sup>6,41b</sup> and Asn404<sup>7,61b</sup> were changed to alanine. These residues were selected as they participate in polar interactions with the ligand and thus were considered more likely to have a measurable effect on ligand binding (Fig. 2a). In addition, these residues interact with the amide and carboxyl functions of MK-0893, which have been demonstrated to be critical for the antagonist activity of the ligand<sup>4</sup>. Fluorescence-activated cell sorting (FACS) analysis using an antibody to the extracellular surface of receptor was performed to compare cell surface expression of mutants with the wild-type receptor. None of the mutations caused any reduction in the cell surface expression levels of the receptor nor had any effect on the binding affinity of glucagon peptide for the orthosteric site (Extended Data Fig. 6 and

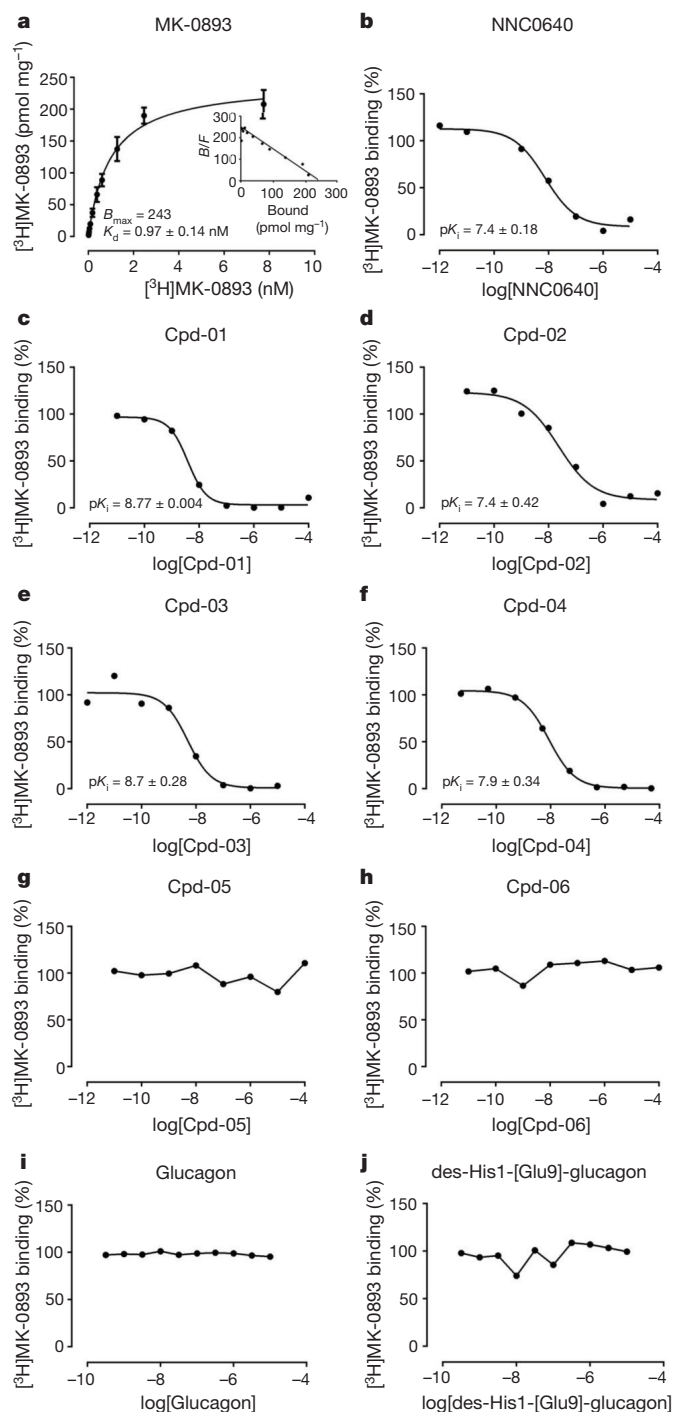




**Figure 2 | Confirming the MK-0893-binding site in GCGR.** **a**, Diagram of ligand interactions in the MK-0893-binding site. Hydrogen bonds are depicted as dashed red lines with distances between heavy atoms in Å. GCGR in ribbon representation is coloured blue, MK-0893 in stick representation is coloured as per Fig. 1. **b**, Comparison of  $pK_d$  of wild type (WT) with the mutants. Data are average of three independent experiments and error bars represent standard error of the mean (s.e.m.).  $P$  values are calculated from a two-tailed  $t$ -test. NS, not significant. The data set for R346E and N404A did not fit the one-site binding unambiguously due to near complete loss of specific binding. ND, not determined.

Extended Data Table 2). Mutation of Arg346<sup>6.37b</sup> to glutamic acid or Asn404<sup>7.61b</sup> to alanine reduced binding levels to close to undetectable above non-specific binding, while mutation of Lys349<sup>6.40b</sup> to alanine or methionine, and mutation of Arg346<sup>6.37b</sup> to alanine, significantly reduced the binding affinity of [<sup>3</sup>H]MK-0893 (Fig. 2b and Extended Data Fig. 7). Consistently, these mutations also reduced the ability of MK-0893 to antagonize the glucagon-mediated increase in cAMP (Extended Data Fig. 8).

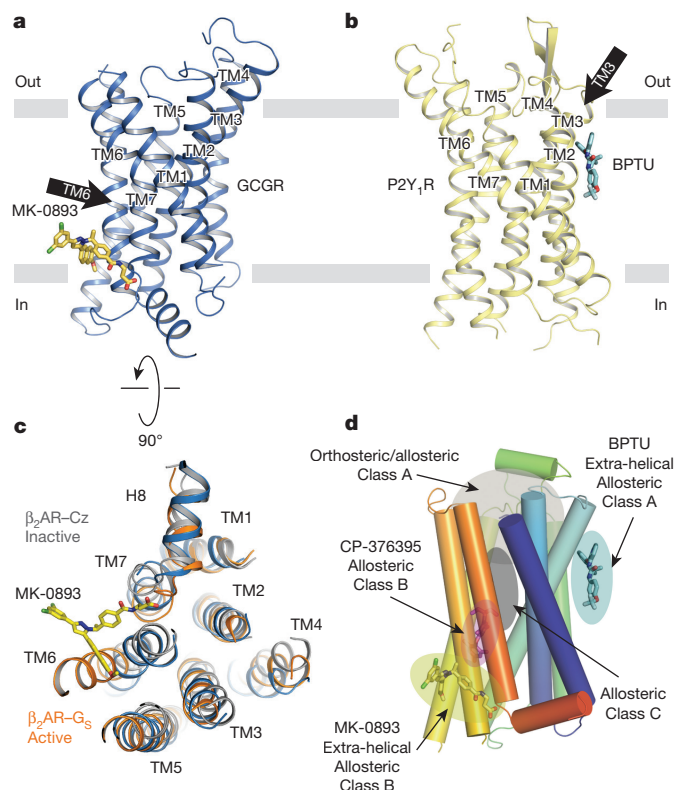
The MK-0893 binding mode suggests that the ligand acts as a clamp holding TM6 in the inactive state and hampering receptor conformational changes required for G-protein coupling (Fig. 4a, c). In the case of class A receptors, activation results in a rotation of TM6 in response to agonist binding, which is transmitted through a rigid-body movement amplified along TM6, thereby altering the interface between TM5 and TM6 and leading to an outward movement of the cytoplasmic end of TM6 to enable G-protein binding<sup>12</sup>. So far, no active structures of class B receptors have been reported; however, the ability of MK-0893 to block signalling of GCGR points to a similar critical role of TM5 and TM6 in class B receptor activation. This is consistent with the



**Figure 3 | Pharmacology of MK-0893.** **a**, Saturation binding of [<sup>3</sup>H]MK-0893 to membranes containing wild-type GCGR. Inset shows the Scatchard plot. **b–j**, Representative competitive binding data for inhibition of [<sup>3</sup>H]MK-0893 binding to membranes containing wild-type GCGR in the presence of indicated compounds (for chemical structures, see Extended Data Fig. 3). Values represent an average of at least three independent experiments  $\pm$  s.e.m.  $K_d$ , dissociation constant.  $pK_i$ , negative logarithm of the inhibition constant.  $B_{max}$ , total concentration of receptors.  $B$  and  $F$  in the Scatchard plot denote bound and free ligand concentrations, respectively.

observation that, despite divergence in the extracellular arrangement of TM helices between class A and B receptors, very good structural conservation exists on the intracellular sides in the inactive conformation<sup>5,11</sup>. An extra-helical binding site was recently described for 1-(2-(2-(*tert*-butyl)phenoxy)pyridin-3-yl)-3-(4-(trifluoromethoxy)





**Figure 4 | MK-0893 allosteric mechanism of action.** **a, b,** Ribbon representation of GCGR (blue) and P2Y<sub>1</sub> receptor (P2Y<sub>1</sub>R; yellow) respectively, viewed parallel to the membrane. MK-0893 in stick representation is coloured as per Fig. 1. BPTU is shown in stick representation with carbon, nitrogen, oxygen and fluorine atoms coloured yellow, blue, red and grey respectively. Potential restrictions on TM movements are indicated. **c,** View of cytoplasmic side of GCGR (rotated 90° from **a**) superposed with the  $\beta_2$ -adrenoceptor ( $\beta_2$ -AR) in complex with carazolol (grey) (Protein Data Bank (PDB) accession 3NY9) and the  $\beta_2$ -AR-Gs complex (orange) (PDB accession 3SN6). **d,** Schematic overview of known binding positions of class A, B and C GPCR ligands.

phenyl)urea (BPTU), an allosteric antagonist of the P2Y<sub>1</sub> receptor<sup>13</sup>, although in this case the binding site was located between TM1, 2 and 3 (Fig. 4b). It is likely that BPTU inhibits the movement of TM3 that is also critical in receptor activation<sup>14</sup>. Together, these structures demonstrate that modulation of helical movements from membrane-proximal surfaces represents an alternative way of modifying the activity of GPCRs.

The identification of an allosteric binding site in the class B GCGR located outside the canonical helical bundle further adds to the diversity of interactions now known to occur between GPCRs and their ligands in modifying receptor activation states (Fig. 4d). The GCGR–MK-0893 structure provides insight into the activation mechanism of class B receptors, as well as facilitating the application of structure-based drug design strategies to discover compounds with improved qualities. Strong conservation of this binding site across human class B receptors indicates that the structural information provided here can be applied to other members of this medically relevant family of GPCRs. In future, drug design paradigms must consider that, in addition to the orthosteric binding pocket, extra-helical binding sites accessed from

the membrane or within the cell may provide alternative strategies to modulate receptor function.

**Online Content** Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 6 August 2015; accepted 9 February 2016.

Published online 25 April; corrected online 11 May 2016

(see full-text HTML version for details).

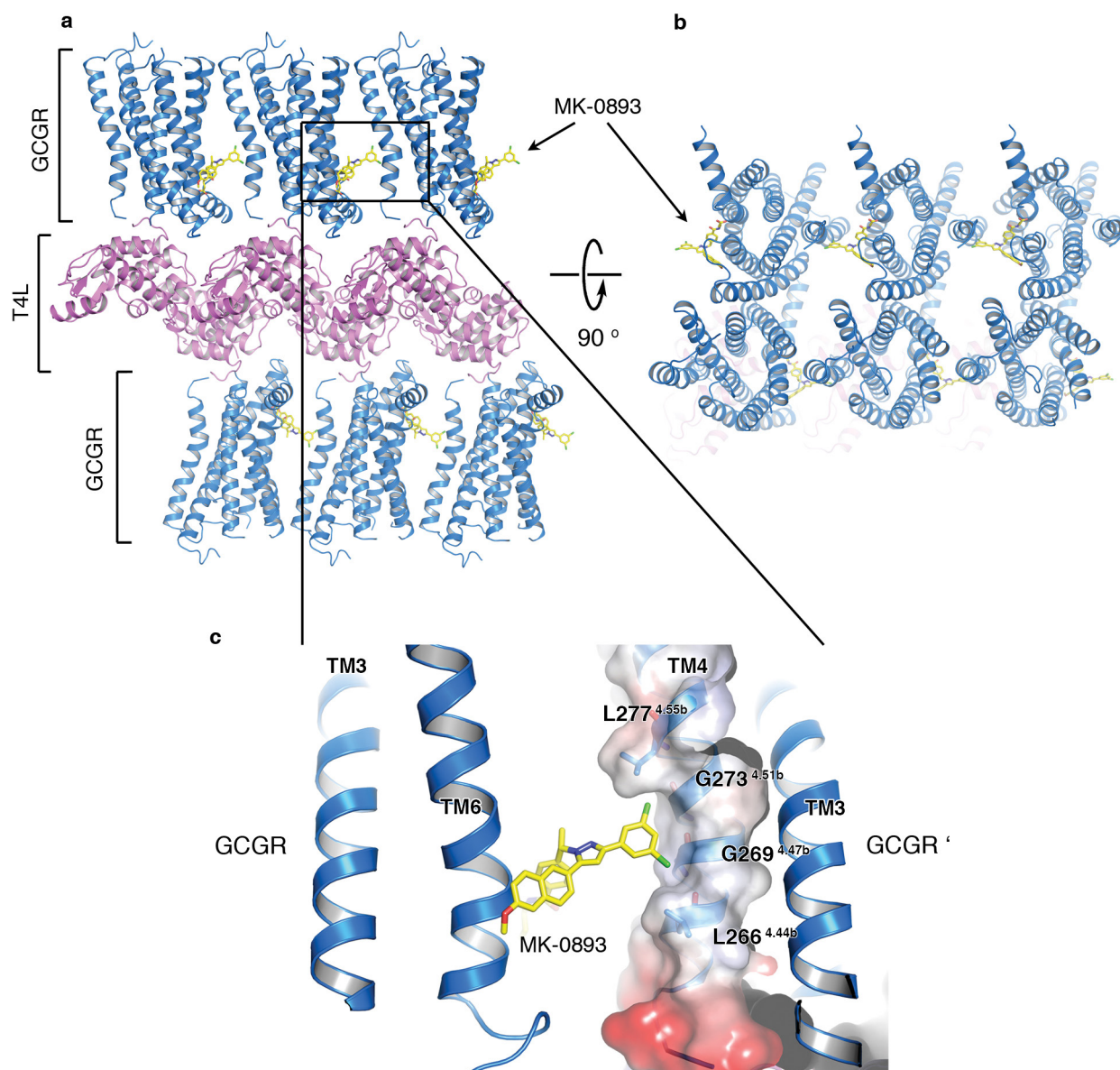
- Ahrén, B. Glucagon—early breakthroughs and recent discoveries. *Peptides* **67**, 74–81 (2015).
- Bagger, J. I., Knop, F. K., Holst, J. J. & Vilsbøll, T. Glucagon antagonism as a potential therapeutic target in type 2 diabetes. *Diabetes Obes. Metab.* **13**, 965–971 (2011).
- Siu, F. Y. *et al.* Structure of the human glucagon class B G-protein-coupled receptor. *Nature* **499**, 444–449 (2013).
- Xiong, Y. *et al.* Discovery of a novel glucagon receptor antagonist *N*-[(4-(1S)-1-[3-(3, 5-dichlorophenyl)-5-(6-methoxynaphthalen-2-yl)-1*H*-pyrazol-1-yl]ethylphenyl)carbonyl]- $\beta$ -alanine (MK-0893) for the treatment of type II diabetes. *J. Med. Chem.* **55**, 6137–6148 (2012).
- Hollenstein, K. *et al.* Structure of class B GPCR corticotropin-releasing factor receptor 1. *Nature* **499**, 438–443 (2013).
- Serrano-Vega, M. J., Magnani, F., Shibata, Y. & Tate, C. G. Conformational thermostabilization of the  $\beta_1$ -adrenergic receptor in a detergent-resistant form. *Proc. Natl Acad. Sci. USA* **105**, 877–882 (2008).
- Shibata, Y. *et al.* Thermostabilization of the neurotensin receptor NTS1. *J. Mol. Biol.* **390**, 262–277 (2009).
- Robertson, N. *et al.* The properties of thermostabilised G protein-coupled receptors (StaRs) and their use in drug discovery. *Neuropharmacology* **60**, 36–44 (2011).
- Ballesteros, J. A. & Weinstein, H. Integrated methods for the construction of three-dimensional models and computational probing of structure-function relations in G protein-coupled receptors. *Methods Neurosci.* **25**, 366–428 (1995).
- Wootton, D., Simms, J., Miller, L. J., Christopoulos, A. & Sexton, P. M. Polar transmembrane interactions drive formation of ligand-specific and signal pathway-biased family B G protein-coupled receptor conformations. *Proc. Natl Acad. Sci. USA* **110**, 5211–5216 (2013).
- Hollenstein, K. *et al.* Insights into the structure of class B GPCRs. *Trends Pharmacol. Sci.* **35**, 12–22 (2014).
- Deupi, X. & Standfuss, J. Structural insights into agonist-induced activation of G-protein-coupled receptors. *Curr. Opin. Struct. Biol.* **21**, 541–551 (2011).
- Zhang, D. *et al.* Two disparate ligand-binding sites in the human P2Y<sub>1</sub> receptor. *Nature* **520**, 317–321 (2015).
- Tehan, B. G., Bortolato, A., Blaney, F. E., Weir, M. P. & Mason, J. S. Unifying family A GPCR theories of activation. *Pharmacol. Ther.* **143**, 51–60 (2014).

**Supplementary Information** is available in the online version of the paper.

**Acknowledgements** We thank D. Hall and J. Waterman at I04, Diamond Light Source, Oxford, UK for technical support. We thank colleagues at Heptares Therapeutics Ltd for suggestions and comments, specifically R. K. Y. Cheng.

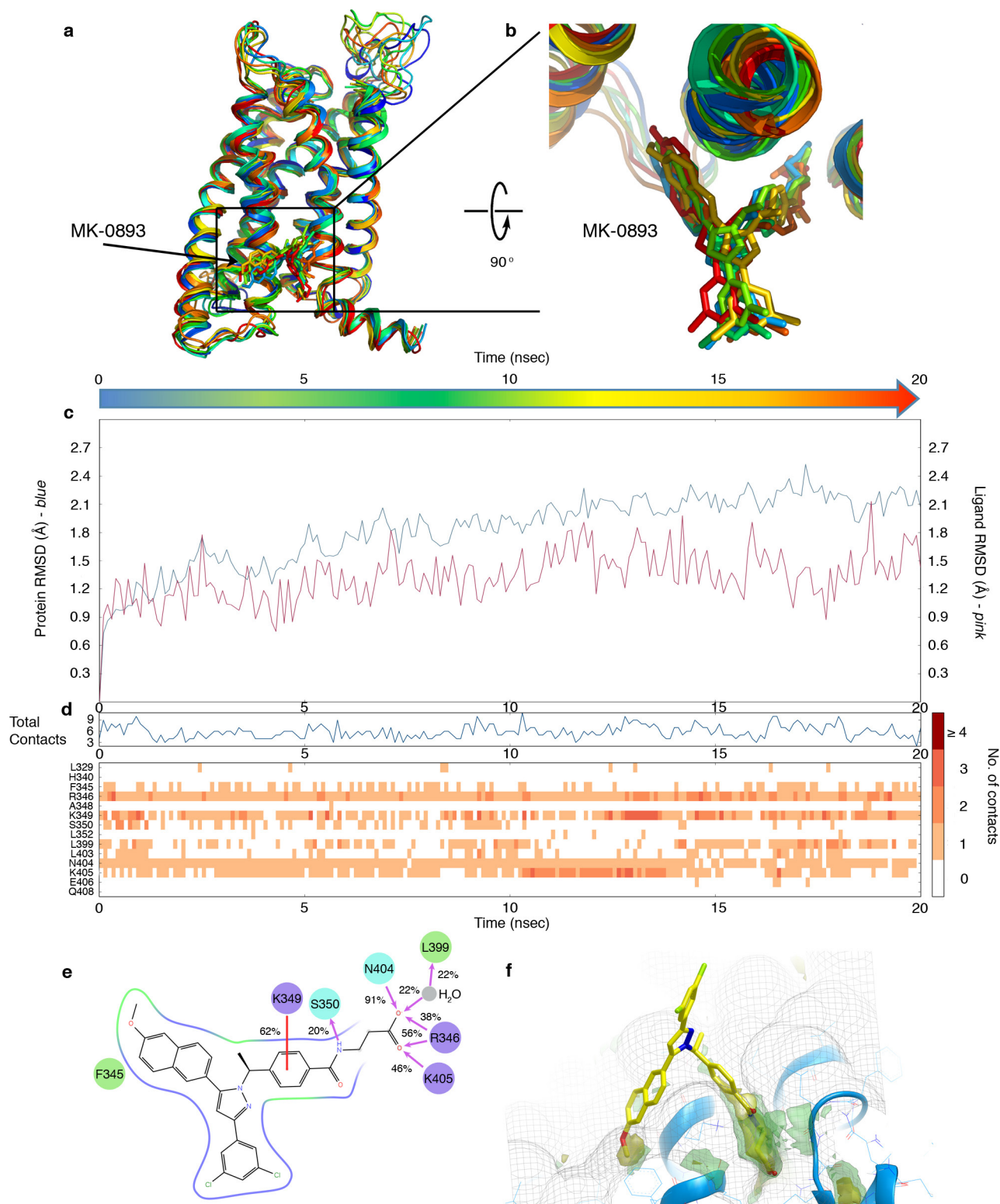
**Author Contributions** D.L. and A.J. devised the strategy and carried out the conformational thermostabilization of the receptor and construct engineering. H.K. established procedures for, and H.K. and S.M.S. carried out expression and optimized purification of the final construct. H.K., S.M.S. and A.S.D. established the platform/protocols for LCP crystallization, harvested crystals, collected and processed X-ray diffraction data, and solved and refined the structure. N.J.R. supported thermostabilization. M.K. and J.C.E. supported expression and purification of the final StaR. A.H.B., I.T. and A.J.H.B. carried out and analysed the pharmacology data. Computational analysis of the structure and modelling was carried out by A.B. S.P.A. identified and sourced the chemical compounds used in the study. Project management was carried out by A.J., R.M.C., M.W. and F.H.M. The manuscript was prepared by A.S.D., A.J. and F.H.M. All authors contributed to the final editing and approval of the manuscript.

**Author Information** Co-ordinates and structure factors have been deposited in the Protein Data Bank under accession number 5EE7. Reprints and permissions information is available at [www.nature.com/reprints](http://www.nature.com/reprints). The authors declare competing financial interests: details are available in the online version of the paper. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to F.H.M. ([fiona.marshall@heptares.com](mailto:fiona.marshall@heptares.com)).



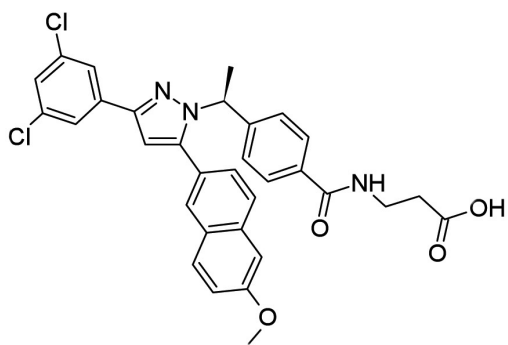
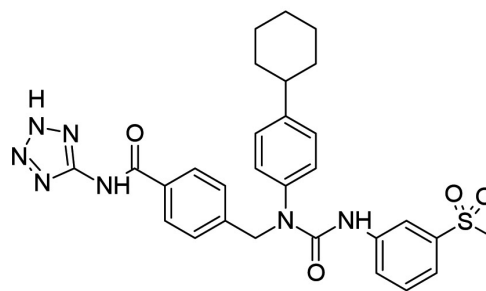
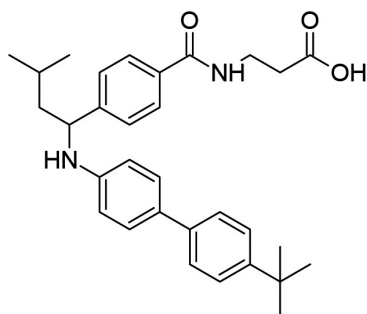
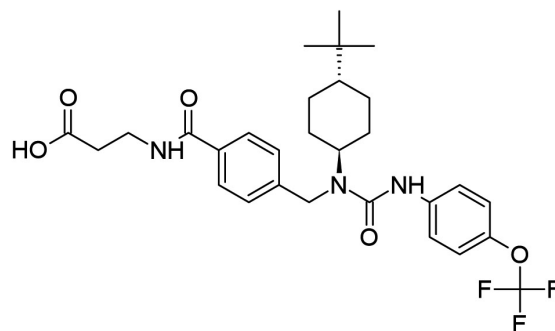
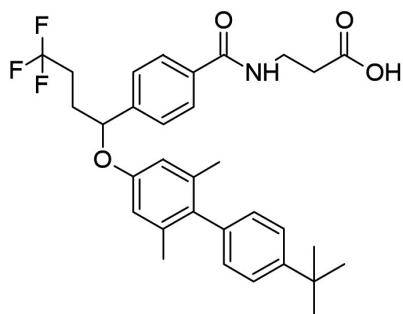
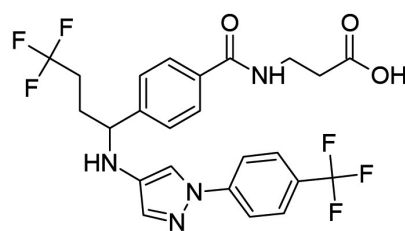
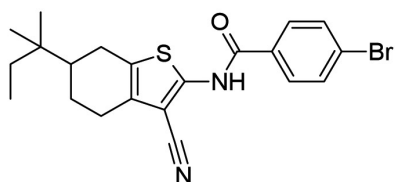
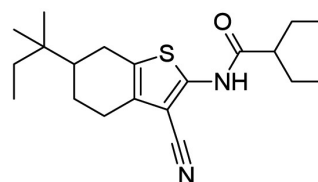
**Extended Data Figure 1 | Packing interactions in the GCGR-StaR(136–417)-T4L primitive orthorhombic crystal system.** **a**, View along *b*-axis, GCGR-StaR TMD in blue ribbon representation, T4L in magenta ribbon representation, MK-0893 in stick representation with carbon, nitrogen, oxygen and chlorine atoms coloured yellow, blue, red

and green, respectively. **b**, View as in **a** rotated 90° to view along *c*-axis. **c**, Close-up of the hydrophobic/shape complementarity interaction of the dichlorophenyl 'head' group of MK-0893 with residues on TM4 of a symmetry-related copy.



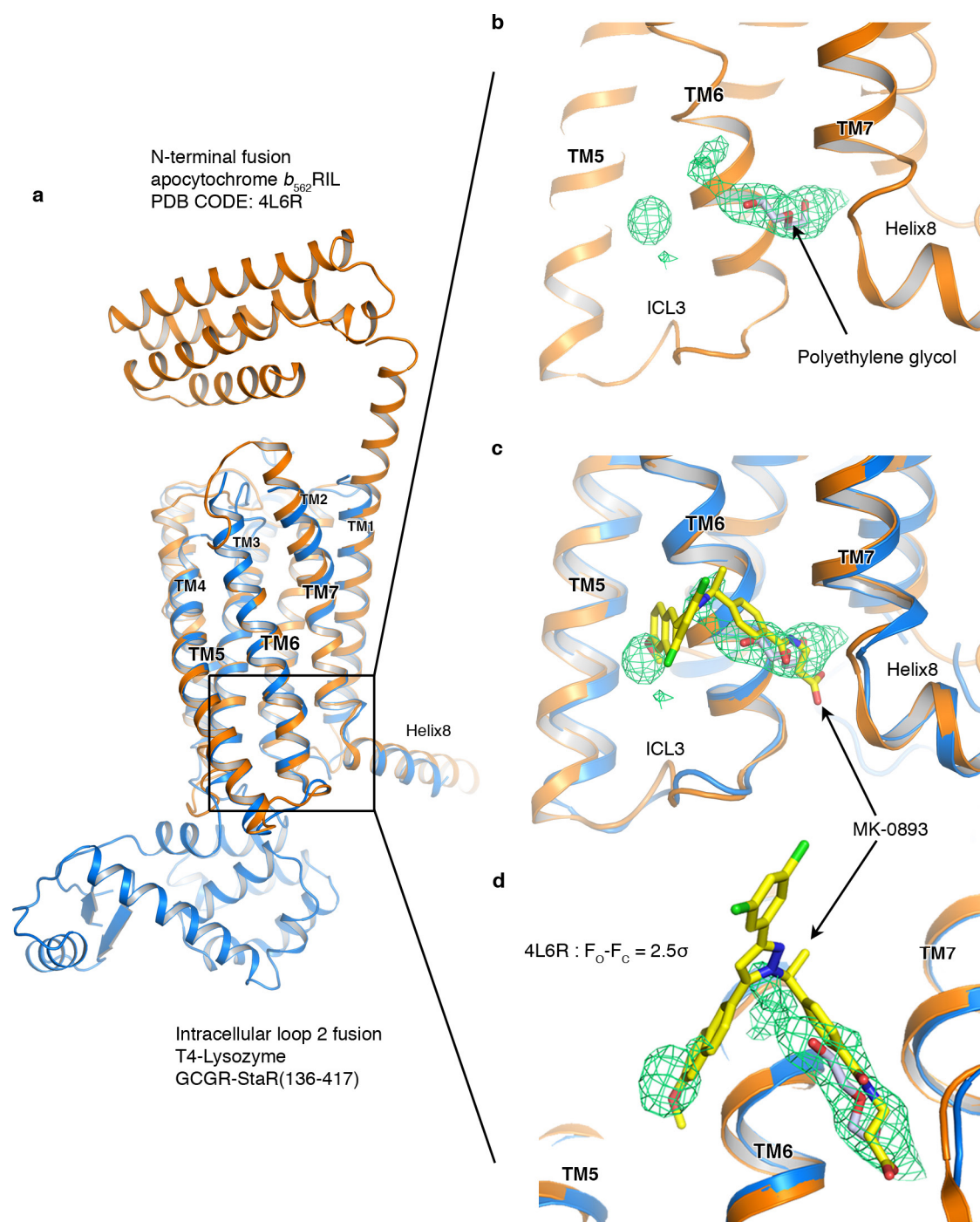
**Extended Data Figure 2 | Molecular dynamics analysis of the GCGR-MK-0893 complex.** **a**, Structural alignment of the wild-type GCGR-MK-0893 complex at 0, 4, 8, 12, 16 and 20 ns molecular dynamics, colour-coded from blue to red as indicated by the arrow. Protein is shown as ribbon, MK-0893 as sticks. **b**, Extracellular view of the ligand conformations during the molecular dynamics simulation, colour-coded as in **a**. **c**, Root-mean-squared deviation (r.m.s.d.) in Å for protein Cα (blue) and ligand heavy atom (pink) during the simulation, after structural alignment to the initial model. **d**, Number of protein-ligand contacts during the simulation: top, fluctuation of the total number of contacts over time; bottom, individual residues interacting with the ligand at a particular time are shown as rectangles colour-coded based on the number of contacts, from white (no contacts) to dark red (four contacts).

**e**, Two-dimensional representation of the ligand-protein contacts. Green circles represent hydrophobic, cyan represents polar and purple are charged residues. Interactions that occur more than 20% of the simulation time in the selected trajectory are shown and the percentage frequency is marked. Hydrogen bonds are shown as pink arrows. A  $\pi$ -cation interaction is shown as a red line. The part of the ligand buried within the membrane is surrounded by a continuous line, while the water-exposed atoms (corresponding to the propionic acid moiety) are indicated by grey circles. **f**, GRID (Molecular Discovery) analysis of the ligand-binding site. The shape of the pocket is shown as grey mesh, hydrophilic and hydrophobic hotspots are shown using green and yellow transparent surfaces, respectively. This panel was prepared in Vida (OpenEye).

**MK-0893****NNC0640****Cpd-01****Cpd-02****Cpd-03****Cpd-04****Cpd-05****Cpd-06**

Extended Data Figure 3 | Chemical structures of compounds analysed in this study.





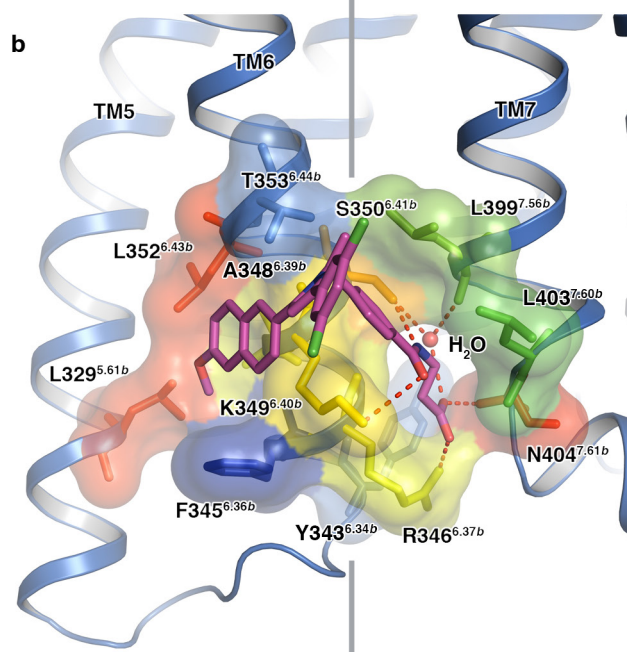
#### Extended Data Figure 4 | Structural superposition of GCGR structures.

**a**, Structural superposition of the GCGR-StaR(136–417)–MK-0893 structure and GCGR–NNC0640 structure (PDB accession 4L6R) in blue and orange ribbon representation, respectively. **b**, Close-up view of the intracellular ends of TM5, TM6 and TM7 in 4L6R, as viewed from within the membrane. A PEG molecule modelled in the 4L6R coordinates is shown in stick representation with carbon coloured grey

and oxygen coloured red.  $m|F_o| - \Delta|F_c|$  map calculated using the 4L6R structure factors and 4L6R coordinates with the PEG molecule omitted, difference density is rendered at  $2.5\sigma$ . **c**, View as in **b**, with the GCGR-StaR(136–417)–MK-0893 structure overlaid with MK-0893 in stick representation with carbon, nitrogen, oxygen and chlorine atoms coloured yellow, blue, red and green, respectively. **d**, Representation as in **c**, close-up and tilted towards the view from the cytoplasm.

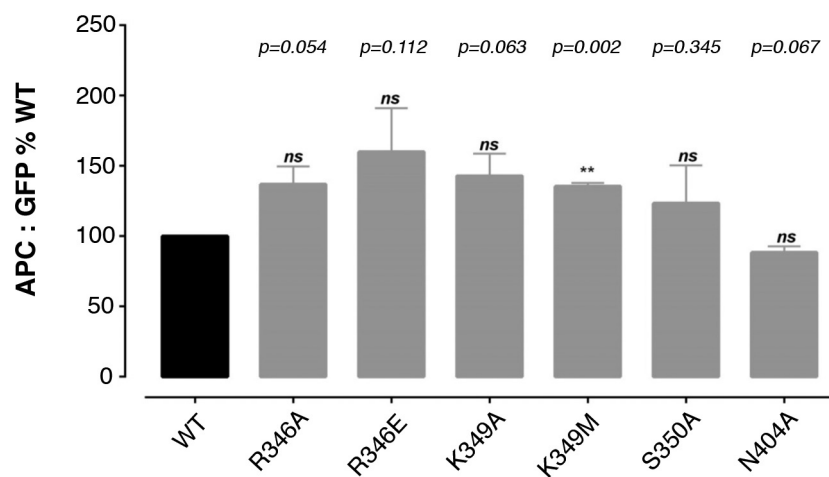
a

	5.61	6.36	6.39	6.43	6.44	6.34	6.37	6.40	6.41	7.56	7.61	7.60
GCGR (P47871)	<b>L329</b>	<b>F345</b>	<b>A348</b>	<b>L352</b>	<b>T353</b>	<b>Y343</b>	<b>R346</b>	<b>K349</b>	<b>S350</b>	<b>L399</b>	<b>N404</b>	<b>L403</b>
CRF <sub>1</sub> R (P34998)	L294	R310	V313	L317	V318	Q308	K311	K314	A315	F362	N367	L366
CRF <sub>2</sub> R (Q13324)	L290	R306	V309	L313	V314	Q304	K307	K310	A311	F358	N363	F362
CALCR (P30988)	L339	L355	V358	M362	I363	M353	K356	K359	A360	I406	N411	C410
CALRL (Q16602)	L316	M332	V335	L339	I340	L330	K333	R336	A337	I383	N388	F387
GIPR (P48546)	L321	L337	A340	L344	T345	Y335	R338	R341	S342	L391	N396	I395
GLP1R (P43220)	V331	C347	A350	L354	T355	I345	R348	K351	S352	L401	N406	V405
GLP2R (O95838)	L365	Y381	A384	L388	V389	Y379	R382	K385	S386	Q435	N440	A439
PACR (P41586)	L331	L349	A352	L356	L357	I347	R350	R353	S354	L399	N404	L403
VIPR1 (P32241)	L319	S337	A340	L344	L345	P335	R338	R341	S342	L387	N392	L391
VIPR2 (P41587)	L306	K324	A327	L331	L332	Q322	R325	K328	S329	L374	N379	L378
SCTR (P47872)	L320	K338	A341	L345	L346	H336	R339	R342	S343	L387	N392	L391
GHRHR (Q02643)	L307	W325	S328	L332	F333	Q323	R326	K329	S330	L375	N380	L379
PTH1R (Q03431)	L385	R404	L407	L411	V412	Q402	K405	K408	S409	I458	N463	C462
PTH2R (P49190)	L340	R359	A362	L366	V367	Q357	K360	K363	S364	I412	N417	C416



**Extended Data Figure 5 | Conservation of residues in the MK-0893 bipartite allosteric pocket.** **a**, Sequence alignment of 15 human GPCR class B members across residues constituting the MK-0893 bipartite binding site. UniProt accession numbers are given. **b**, The GCGR MK-0893 allosteric binding site is shown in surface representation with

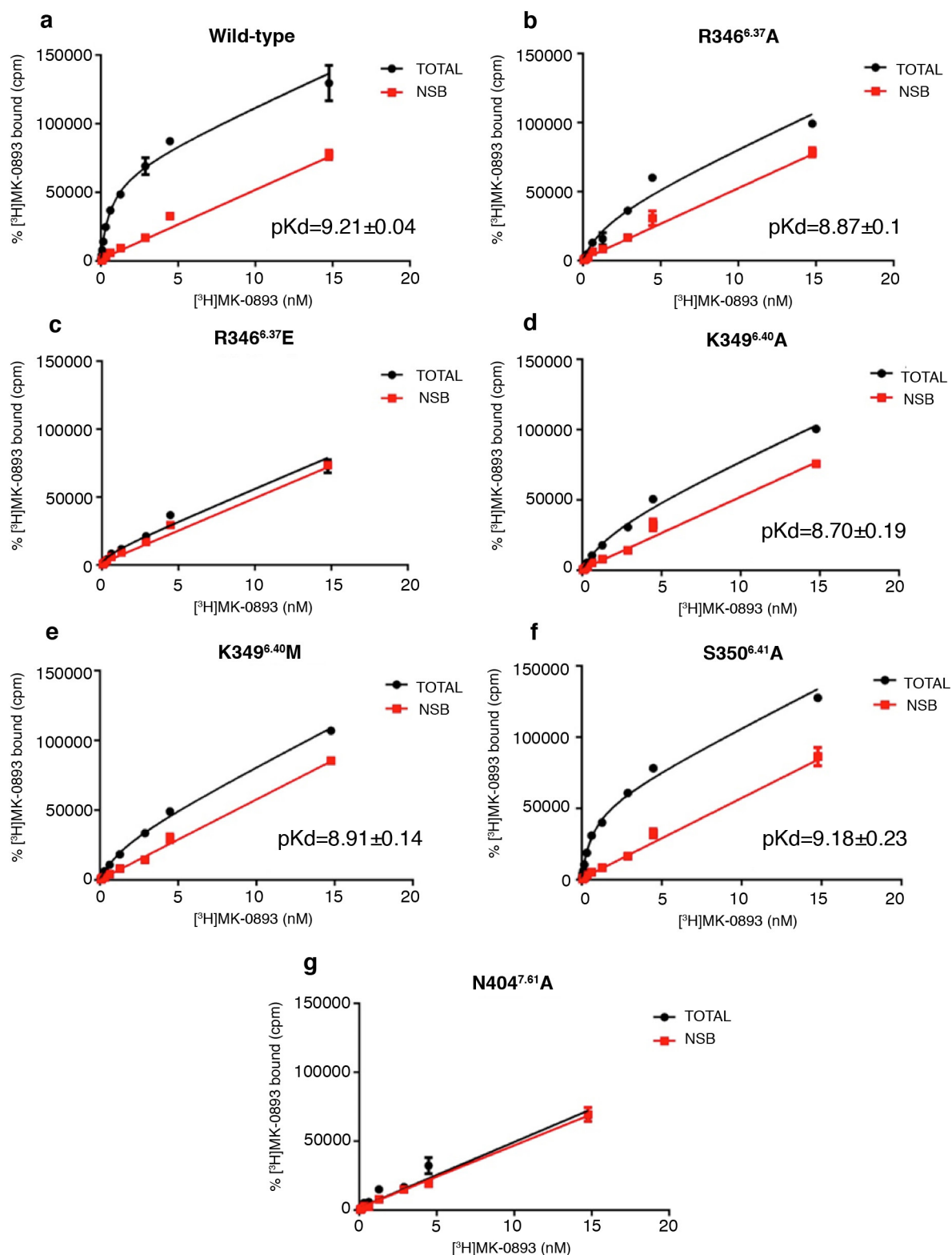
residues in **a** coloured in rainbow spectrum according to conservation level (red = 100%; blue = 0%). MK-0893 is in stick representation with carbon, nitrogen, oxygen and chlorine atoms coloured purple, blue, red and green, respectively.



### Constructs

**Extended Data Figure 6 | Expression of GCGR mutants.** Cell-surface expression of GCGR mutants determined using FACS. Data are expressed as the ratio of APC (cell surface expression) to GFP (total expression)

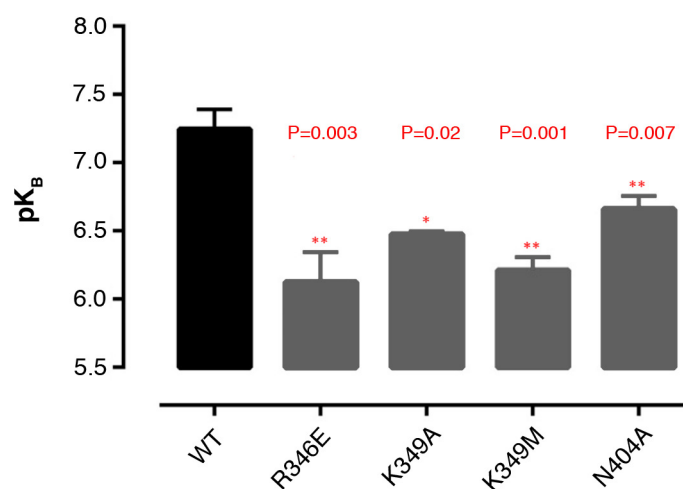
and calculated as percentage of wild type. Experiments were carried out in triplicate and error bars indicate s.e.m. *P* values are derived from an unpaired two-tailed *t*-test.



**Extended Data Figure 7 | Saturation binding analysis of mutants with [<sup>3</sup>H]MK-0893.** a–g, Saturation binding of [<sup>3</sup>H]MK-0893 to membrane containing the indicated variants of GCGR. Data are representative of three independent experiments. pK<sub>d</sub> values are average of three

independent experiments and error bars represent s.e.m. *P* values are calculated from a two-tailed *t*-test. The data set for R346E and N404A did not fit the one-site binding unambiguously due to near complete loss of specific binding.





### Constructs

**Extended Data Figure 8 | Effect of binding-site mutations in functional assay.** Cells expressing either wild type or the indicated mutants were stimulated with a concentration range of glucagon in the presence of increasing concentrations of MK-0893. After determination of levels of cAMP generated, the data were analysed by global fitting to the

half-maximum effective concentration ( $EC_{50}$ ) allosteric shift equation.  $pK_B$  is the negative log of the antagonist affinity that mediates the allosteric inhibition of the glucagon response. Data are an average of three independent experiments and error bars represent s.e.m.  $P$  values are derived from an unpaired two-tailed  $t$ -test.

**Extended Data Table 1 | Data collection and refinement statistics for GCGR-StaR(136–417)–T4L–MK-0893**

<b>Data collection</b>	
Number of crystals	9
Space group	P2 <sub>1</sub> 2 <sub>1</sub> 2 <sub>1</sub>
Cell dimensions	
a, b, c (Å)	37.6, 71.5, 183.1
$\alpha$ , $\beta$ , $\gamma$ (°)	90.0, 90.0, 90.0
Number of reflections measured	52,568
Number of unique reflections	16,065
Resolution (Å)	32.73 - 2.50 (2.60 - 2.50)
R <sub>merge</sub>	0.157 (0.768)
CC <sub>1/2</sub> **	0.984 (0.469)
Mean I/sd(I)	6.0 (1.6)
Completeness (%)	91.1 (90.2)
Redundancy	3.3 (3.3)
<b>Refinement</b>	
Resolution (Å)	19.97 - 2.50
Number of reflections (test set)	15,222 (777)
R <sub>work</sub> /R <sub>free</sub>	0.226 / 0.263
Number of atoms	
All	3,642
Protein	3,353
Ligand	41
Others (Lipids, ions, waters)	248
Average B factors (Å <sup>2</sup> )	
All	66.7
GCGR	47.8
T4L lysozyme	101.2
Ligand	37.2
Others (Lipid, ion, water)	51.0
RMSD	
Bond lengths (Å)	0.0038
Bond angles (°)	0.799
Ramachandran statistics	
Favored regions (%)	97.8
Allowed regions (%)	2.2
Outliers (%)	0.0
<i>MolProbity</i> overall score (percentile)	1.25 (100th percentile)

\*Values in parenthesis indicate highest resolution shell.

\*\* CC<sub>1/2</sub> - see Diedrichs & Karplus, *Acta. Cryst.* (2013). D69, 1215-1222.

Extended Data Table 2 | Further pharmacological characterizations of different GCGR constructs

**a**

	GCGR-WT	GCGR-StaR (136-417)	GCGR-StaR (136-417)-T4L
<b>MK-0893</b>	9.04 (0.52)	9.23 (0.57)	9.01 (0.47)
<b>Cpd-03</b>	8.82 (0.24)	8.78 (0.19)	8.52 (0.11)

**b**

GCGR-WT	R346A	R346E	K349A	K349M	S350A	N404A
8.9 (0.01)	8.86 (0.23)	8.78 (0.04)	8.84 (0.11)	9.15 (0.01)	9.00 (0.02)	8.86 (0.01)

**a.** Analysis of MK-0893 and Cpd-03 binding to different GCGR constructs.  $pK_i$  values are calculated from competition studies with [ $^3$ H]MK-0893 using membranes isolated from HEK293T cells transiently expressing the indicated constructs. Data are shown as the mean of three independent experiments with standard deviation displayed in parentheses. **b.** Analysis of glucagon binding to wild type and allosteric binding site mutants.  $pK_d$  values are calculated from homologous competition studies with [ $^{125}$ I]glucagon using membranes isolated from HEK293T cells transiently expressing the indicated constructs. Data are shown as the mean of two independent experiments with standard deviation displayed in parentheses.

CORRIGENDUM

doi:10.1038/nature16543

Corrigendum: Essential roles of PI(3)K–p110β in cell growth, metabolism and tumorigenesis

Shidong Jia, Zhenning Liu, Sen Zhang, Pixu Liu, Lei Zhang, Sang Hyun Lee, Jing Zhang, Sabina Signoretti, Massimo Loda, Thomas M. Roberts & Jean J. Zhao

Nature 454, 776–779 (2008); doi:10.1038/nature07091

In Fig. 3b of this Letter we inadvertently used the wrong images (partial duplicates of images representing *p110α*<sup>−/−</sup> mice) in the panels representing *p110β*<sup>−/−</sup> mice. The corrected Fig. 3b is shown in Fig. 1 of this Corrigendum. The conclusions of the paper are not affected; knocking out either *p110α* or *p110β* does not affect the growth of the normal prostate.

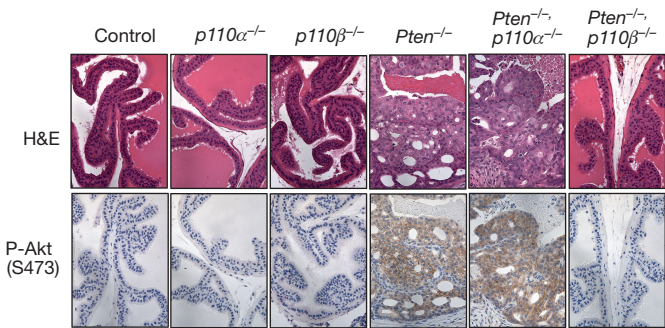


Figure 1 | This figure shows the corrected Fig. 3b of the original Letter.



## CORRIGENDUM

doi:10.1038/nature16999

### Corrigendum: Structural and functional features of central nervous system lymphatic vessels

Antoine Louveau, Igor Smirnov, Timothy J. Keyes, Jacob D. Eccles, Sherin J. Rouhani, J. David Peske, Noel C. Derecki, David Castle, James W. Mandell, Kevin S. Lee, Tajie H. Harris & Jonathan Kipnis

*Nature* **523**, 337–341 (2015); doi:10.1038/nature14432

We would like to correct this Letter, which demonstrated the molecular characteristics and functional nature of lymphatics serving the central nervous system (CNS), by adding two reference citations, of which we became aware after publication. Foldi *et al.*<sup>1</sup> postulated a CNS–lymphatic link, although their findings described lymphatic vessels located at the base of the skull, which differs from the location of the vessels that we identified and functionally characterized. Andres *et al.*<sup>2</sup> presented incidental histological evidence of dural vessels in rats that, on the basis of ultrastructural features, appeared to be lymphatic. We also wish to clarify that the designation of these lymphatics as ‘CNS lymphatics’ in our Letter refers to their function of CNS/cerebrospinal fluid drainage.

1. Foldi, M. *et al.* New contributions to the anatomical connections of the brain and the lymphatic system. *Acta Anat.* **64**, 498–505 (1966).
2. Andres, K. H., von Düring, M., Muszynski, K. & Schmidt, R. F. Nerve fibres and their terminals of the dura mater encephali of the rat. *Anat. Embryol.* **175**, 289–301 (1987).

# CORRECTIONS & AMENDMENTS

---

## CORRIGENDUM

doi:10.1038/nature17177

### **Corrigendum: Bees prefer foods containing neonicotinoid pesticides**

Sébastien C. Kessler, Erin Jo Tiedeken, Kerry L. Simcock, Sophie Derveau, Jessica Mitchell, Samantha Softley, Amy Radcliffe, Jane C. Stout & Geraldine A. Wright

*Nature* **521**, 74–76 (2015); doi:10.1038/nature14414

In this Letter, the author Amy Radcliffe was acknowledged but erroneously omitted from the author list. We wish to include her in the author list as shown, associated with the Institute of Neuroscience, Newcastle University affiliation. She contributed data to the choice experiments with bumblebees using the pesticide thiamethoxam. This has been corrected in the online versions of the paper.

# CAREERS

**PERSONAL TRIUMPH** Regaining health, gaining a degree **p.281**

**EXPO IN SAN FRANCISCO** Recap of Naturejobs in the 'City by the Bay' [go.nature.com/51ssik](http://go.nature.com/51ssik)

**NATUREJOBS** For the latest career listings and advice [www.naturejobs.com](http://www.naturejobs.com)



## PEER REVIEW

# Close inspection

*To improve your own papers, learn how to evaluate other scientists' work.*

BY QUIRIN SCHIERMEIER

Before she had even defended her doctoral thesis, Brazilian student Rita Santos began to receive requests for her expert opinion. Her work on beak development in octopus larvae — along with her knowledge, care and keen judgement — had left an impression on scientists in the field and early on in her career, she was invited to become a peer reviewer.

Matthias Starck, a zoologist at the Ludwig Maximilian University of Munich in Germany and editor-in-chief of the *Journal of Morphology*, sent an invitation to Santos after receiving a

recommendation from her supervisor. "I was a bit hesitant at first," he says, "but the reports she turns in are just superbly thoughtful and well written."

Peer review is the backbone of modern science, and academic researchers are expected to participate in the endeavour. Although time consuming, delving deeply into someone else's paper can benefit a scientist's own work. The process allows peer reviewers to read about research before it is generally known and to gain insight into how other scientists write manuscripts and present data. "I've learned a lot about science and the process of publishing it," says Santos, who studies marine ecosystems

at the Alfred Wegener Institute of Polar and Marine Sciences in Bremerhaven, Germany. "And you learn how to be critical without being impolite or discouraging to others."

Whether or not they plan to pursue an academic career, junior researchers should get involved in peer review, says Sarah Blackford, a career adviser with the Society for Experimental Biology in London. "Not only will it help you to hone your power of judgement," she says, "but it is also a great way to broaden your knowledge and demonstrate transferable skills for offering an authoritative view to your peers."

## HAND-ME-DOWN PAPERS

Young scientists typically get their start as reviewers through supervisors or lab leaders, who may be overburdened or need to turn to junior team members who are familiar with specific methods or technology. Graduate students generally are not recognized for their ability to conduct independent peer review unless, like Santos, they are already establishing an academic reputation by publishing first-author papers. But they can gain experience by helping their supervisors or senior colleagues to prepare reviews.

"If I am too busy, or a manuscript is a little outside my field, there is nearly always an opportunity to propose postdocs and other early-career researchers who have expertise in the area requested," says Ros Gleadow, a plant physiologist at Monash University in Melbourne, Australia. "They might then get invited by the journal to conduct the review."

Even if they aren't invited, another natural first step is to review a paper jointly with seasoned colleagues or under their mentorship, says Emma Ganley, co-editor-in-chief of the journal *PLoS Biology*. Senior scientists might be better placed to judge a finding's weight and general significance, but junior researchers are often more up to date on methods and technology — proficiencies that any journal editor will appreciate.

"Young reviewers are extremely good in raising technical issues such as those related to microscopy or molecular techniques," says Bernd Pulverer, head of scientific publications at the European Molecular Biology Organization (EMBO) in Heidelberg, Germany. Editors of EMBO journals encourage senior reviewers to involve trusted early-career lab members in peer reviews, provided that they have done experimentation in the relevant field. Their background experience will ►

► help them to carry out the key components of peer review: they must be able to assess whether work is new to the field and original enough to deserve publication — and by the journal in question.

They need to be able to evaluate the quality of data, look for potential inconsistencies and ascertain whether the methods and experiments are appropriate. If they see flaws or holes, they will be expected to suggest that the authors do more analysis or more experiments. And if they think that a paper is incomprehensible or biased (or plain tripe), they are obliged to tell journal editors just that.

Reviewers will assess whether a study is conceptually valid or technically sound, if its arguments are coherent and if claims and conclusions are sufficiently backed up by the data. Many journals will also ask whether the results challenge or confirm established concepts, and if they significantly advance the field at hand.

“Don’t try to dictate to us what we should be publishing, but do provide strong arguments and detailed justifications of any statements you make,” says Karl Ziemelis, chief physical-sciences editor at *Nature*. “Just saying that this or that isn’t a big deal in your field is much too vague. We would like to know why you think so, and how you came to that conclusion.”

## THE STARTING GATE

New reviewers may be uncertain of what they are expected to produce and how overtly critical they should be. “I knew I was to assess the scientific strengths and weaknesses of the manuscript,” says Santos. “But I wasn’t quite sure, at first, how deeply I should go into things like length, structure and language.” If they are at all confused, they should consult a seasoned



Marine ecologist Rita Santos out in the field.

## NUTS AND BOLTS

### Become a peer-review legend

- Formal courses in peer review are rare or absent, so seize the opportunity if lecturers offer exercises in discussing papers. Journal clubs are also a helpful way to gain some experience.
- To become a reviewer, you need to make yourself known. A good way to build up trust with journal editors is to approach them at conferences and meetings and show them your work.
- Once you get a manuscript, read it through once, carefully. Let it settle a day before you proceed.
- Establish whether the science is compatible with the scope of the journal.
- Outline the novelty of the science and judge the significance of the results: how do they advance the field?
- Comment on the quality of the science and validity of the results.
- Ask yourself the following questions:
  - Is the argument logical?
  - Are the methods suitable and results plausible?
  - Are the findings adequately described and discussed?
  - Are the claims and conclusions justified by the data?
  - Is the interpretation of the data appropriate in light of available theory?
  - Have the authors conducted all appropriate controls?
  - Is there adequate replication?
  - Are key papers in the field cited?
- Give an opinion as to whether the paper should be published, revised or rejected.
- Describe any extra experimentation or data analysis needed to warrant publication.
- Ask journal editors for feedback: what was your review like? Was anything missing? **U.S.**

reviewer, or contact the journal editor who commissioned the review, advises Ziemelis. They should also tell the editor if they feel that they might lack competence — or the time — to do a proper review. If the field in question is too distant from their own niche, they may need to decline to review a manuscript, or suggest someone who is more appropriate.

“Do tell editors if you are happy to comment on one aspect of a paper but not on another,” says Pulverer. Journal editors also appreciate it when a researcher recommends colleagues who might be better placed to evaluate a paper or any specific aspects of the science. If a peer reviewer brings in a student or technical specialist to help out, those people should be named as contributing reviewers.

Similarly, Ziemelis says, researchers should tell the journal editor if they think that they are too closely affiliated with an author to judge the science neutrally. Any conflict of interest — personal, financial or owing to direct competition — renders a scientist unsuitable as a reviewer. It is always better to over-declare than to under-declare, says Irene Hames, an independent publishing consultant in York, UK, and former director of an international organization called the Committee on Publication Ethics.

Novice reviewers should also find out whether the journal offers ‘double-blind’ peer review, in which authors can request that their names and affiliations be withheld. A reviewer will need to decide whether she or he is comfortable reviewing the work of an anonymous author. Conversely, in the case of ‘open’ peer review, the author’s and reviewer’s identities are disclosed. But this model offers new reviewers the chance to look at what others have written and how authors have responded to comments.

If a junior researcher is contacted by a journal that they have never heard of, they should be cautious. An invitation from what might be a new or relatively unknown small journal isn’t necessarily a reason to decline, but journals with questionable peer-review and publishing standards are increasing in number. If a journal says that it is open access, researchers should check whether a journal is listed on the Directory of Open Access Journals ([www.doaj.org](http://www.doaj.org)) or the Open Access Scholarly Publishers Association ([www.oaspa.org](http://www.oaspa.org)). They should look for recognized experts on a journal’s editorial board, and contact them to verify credentials and peer-review standards.

## THE WRITE UP

The review itself involves several steps (see ‘Become a peer-review legend’). The first is to plan enough time and to stay in close contact with editors. There is no one-size-fits-all estimate for how long it takes to write a good review, but scientists should expect to spend at least eight hours and up to several weeks, say veteran reviewers.

Sloppy work or unresponsiveness might prompt editors to drop a reviewer — which could mean losing the respect of peers and colleagues and diminishing the chance of being added to editorial boards. It could also taint a researcher’s reputation with editors of journals in which they may want in future to publish their own work.

After an initial general read of the manuscript, novice reviewers should wait a full day or so before getting into the technical details and starting to draft a properly phrased review, says structural biologist Stephen Curry of Imperial College London. “Sit back and think



how you would like a constructive review to be written if you were the author of the paper," he says. Snarkiness or scorn should not be present. "Derisiveness, aggressiveness or rivalry have absolutely no place in a review," Curry adds.

The document should start with a short, cohesive summary of the paper, says Pulverer, followed by comment on experimental design and the validity of controls. A key point of any review of biological work, he says, is whether the data and their interpretation support the reported findings. "We'd like reviewers to outline precisely what extra tests they think are needed and why," he says. Reviewers should also make clear whether more experiments are essential or merely desirable.

The specific technical and editorial advice that reviewers are expected to provide depends largely on the subject area and the scope of a journal. Validating a twist in string theory or cosmology calls for a different approach than reviewing the results of an astronomical observation, geological fieldwork or clinical trials. If asked to assess theoretical work, a reviewer should focus on equations and their interpretation.

Most studies will require reviewers to examine observational and experimental data contained in supplementary material (or external repositories) and their representation in graphs and figures.

Reviewers should check guidelines for authors and reviewers carefully to be sure that they properly understand a journal's scope, how novel and 'big' any science must be to get published there, and whether referee reports and the authors' responses will be published online.

If the latter is the case, as it is for the EMBO journals, scientists should look at other reviews and authors' responses. This is a good way for novice reviewers to get a sense of the appropriate length and structure expected, and of the journal's overall review process, says Hames. If such information is ambiguous or unavailable, they should ask the journal for specifics.

Assessing the work of others nurtures critical thinking in ways that few other ventures can match. But at the end of the day, says Alaa Ibrahim, an astrophysicist with the American University in Cairo, it is good for authors to have others dissect their submitted work. "The worst thing," he says, "is that your science gets published just to be proven faulty or wrong soon after." ■

**Quirin Schiermeier** is a Nature correspondent in Munich, Germany.

# TURNING POINT

## Intelligence programmer

*Computer scientist Damien Anderson overcame a lengthy illness to pursue an award-winning PhD project in artificial-intelligence (AI) research at the University of Strathclyde in Glasgow, UK. After regaining his health, he had to wrestle with a crisis of confidence.*

### What led you to study AI?

I've been interested in computers since my dad bought a video-game console, when I was five. I grew up in a deprived area of Scotland called North Lanarkshire, so it was a big deal at the time. Later, I had serious health issues — undiagnosed pneumonia led to chronic-fatigue syndrome — which left me bed-bound from age 14 to 20. I replaced conventional education with the computer, teaching myself subjects that I was interested in. It was a negative time, but positive things came out of it. It gave me time to learn the things I wanted to learn.

### How did you move forward once you were well?

When I had my strength back, I worked at a call centre fixing computers for four years. I decided that if I could do that, I was now physically able to stick to a degree. My confidence had been zapped by being ill for so long. I wanted a piece of paper that said I was capable of doing more than answering phones.

### Can you describe your journey into university?

The hardest decision I ever made was to go back into education. I didn't have high-school qualifications, and so I had to prove myself. I did a national qualification in the form of an introductory course to digital media, and then completed a two-year diploma at the City of Glasgow College — a gateway to university if you don't have enough qualifications. I focused on software-programming languages. After that, I was allowed to enter the University of Strathclyde as a second-year student.

### Were you intent on doing video-game design?

Early on, yes. But when I got to the university's department of computer and information science, I was really impressed by the people and their projects, which included AI. I decided to do a software-engineering degree. But to be honest, my initial goal was just to get a degree. I approached it as if I just had to survive my time in university. It was a game of attrition, and I would beat it through pure persistence.

### What pushed you to do more?

My undergraduate programme offered an optional placement year in industry. When I looked at the list of places that students had



gone before, CERN, Europe's particle-physics lab near Geneva, stood out. I was determined to do well — not just get through it. I studied harder to get the grades necessary.

### How was your time at CERN?

It was a dream. I expected to be surrounded by Einsteins and Feinmans, but these are normal, determined people like me, which was eye-opening. We were using real-time decision-making processes, called scrum, to develop machine-protection software for the Large Hadron Collider. After seven months there, I was named scrum master — essentially, team facilitator. It made me feel extremely valued. I came back after 14 months and finished a final-year project that won 2 awards, which helped me to secure funding from the Carnegie Trust for the Universities of Scotland to conduct a PhD.

### What are you working on now?

The big hurdle in my field now is building AI systems that are able to solve a variety of problems, including ones they've never seen before. The Google DeepMind team — which just announced that its AI, called AlphaGo, won against the world's top Go player — is also funding a competition to build AIs able to solve more than one problem. I'm working on that. One of the best platforms to carry out that project is in video games, because there are so many types — from puzzles to role-playing to strategy.

### How have you handled the attention that your work has received?

The publicity has at times got me way out of my comfort zone, but it's a great confidence boost. I've decided to say yes to every opportunity. ■

**INTERVIEW BY VIRGINIA GEWIN**

This interview has been edited for length and clarity.

# VENICE, VERSION 9.0

*A tale of multiple cities.*

BY PRESTON GRASSMANN

Shifting through different versions of the city, I settle on a drowned vision of the Venetian streets. I wait for her at a table on the make-shift patio of a sunken cathedral, watching the avatars of tourists pass. She comes in a silk dress that flows behind her like swirls of smoke, her copper hair like fire. The illusion is broken only by the swiftness of her motion. As she sits down, I stare at the curve of her neck, the elegance of an angle I could never quite capture in code.

"A bit sacrilegious, don't you think?" she asks, looking up at the cathedral window, where stained glass frames the dusk with a scene from the last supper.

"In their version of Venice, the servers are acolytes walking the aisles with candles and contribution trays." The waiter appears, placing a glass of wine at the table. "This is a sacrament."

"You've made some changes here." She points up at a virtual vaporetto, the engine sounds muted as if heard from several metres below the surface of the Adriatic. It slows down at the edge of the pier, letting the tele-present tourists disembark. They step out into the sea, transform into their chosen avatars and begin their descent to the sunken city.

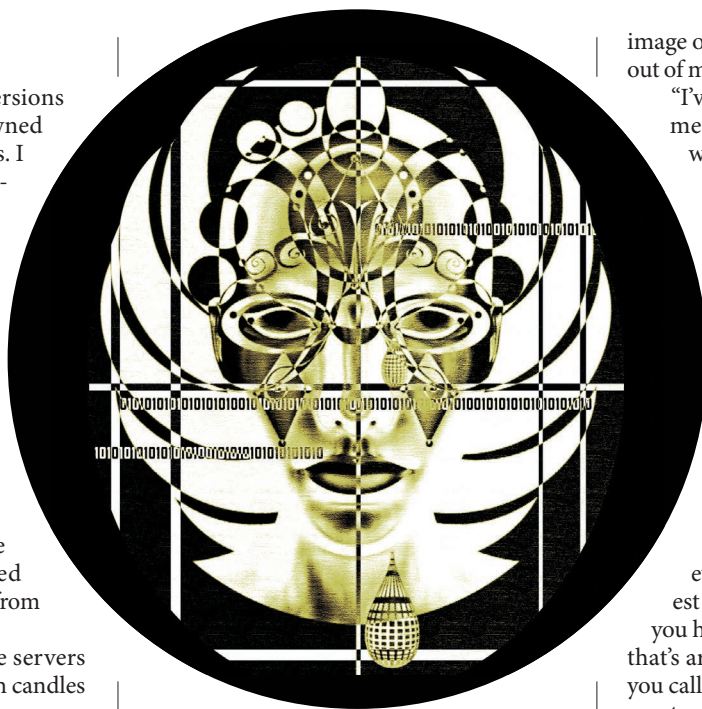
"Thank you for noticing. I've added a few things to make the experience more real."

She rolls her eyes, looking up at the Bridge of Sighs, where a school of exotic fish drifts slowly past the windows. "Do you even know what that means?"

"The wine is real," I say. "I can't say much for the tourists, though."

"This is all a game," she says, leaning forwards, raising the glass in her hands and studying the wine swirling inside. "Faithful algorithms made into virtual-estate, illusions bought and sold on a phantom market."

"Do you really think the pre-virtual centuries were all that different? None of these monuments would have been built without games of power and influence. Do you know how many versions of Venice there were before the virtual cities came along? The 'Queen of the Adriatic', 'City of Water', 'City of Masks', 'City of Bridges', 'The Floating



City' ... Shall I go on?"

"Why did you bring me here?"

"What if I could bring back the old Venice, the one from your own childhood?"

She stops swirling her glass and stares at me over the rim. "What are you talking about?"

The scene fades from view, colours brightening over the streets and buildings, resolving into another version of the city. Sitting where we are, outside the cathedral, I let the new vision open slowly. This is the Venice of her memories, where she often walked with her mother when she was a child, the old storefronts with their handmade signs — shops that sold silver-plated tableware, perfumes, women's hats and ornate glass, gleaming brightly behind the windows. She recognizes the name on the sign near the cathedral, the canal where she once lived.

"No one inhabits this version yet, so the people you see here are default avatars."

"You've turned my memory into an algorithm," she says.

"I thought you would've liked this. This is the only version that will ever exist, outside of your own memory."

➔ **NATURE.COM**  
Follow Futures:  
@NatureFutures  
f go.nature.com/mtoodm

"But it's not real," she says, tapping the table with her finger. A metal sound clanged beneath the

image of the wood. "You've assembled this out of my head."

"I've reconstructed part of it from your memories. Other parts I've filled in with recordings, pictures and public records."

"I don't want to live in my own memories. You should have left them where they were."

"When you told me about your memories of Venice, you said that you wished you could be back in that time. I thought you would've liked this ..."

"It's not just about the time, but the people who inhabit that space. What you've created is only a simulation. It's not real."

"Tell me, what is reality? Your eyes filter everything but the narrowest range of the light. Everything that you hear is only a fragment of the sound that's around you at each moment. What you call 'real' comes from a perception that constantly changes, altered by chemistry and memory."

"Look," she says, reaching out to pass her hand through me. "This is what I know is real. I can't feel you. I can't touch you. There is nothing to hold on to." She stands up slowly. She brushes a hand across her cheeks. The tears are like glass at the edge of her eyes, shaped by forces I can't understand, insoluble in the virtual sea. "I need to be out there in the world, even if it is broken."

I watch her walk away, her copper hair drifting in the wind of another Venice, the nape of her neck a question in the coded landscape of her memory. The only answer I have is that I am also an artefact assembled from memory, an algorithm made out of silicon instead of biology.

I nod to the waiter, placing the virtual money on the table. In another version of the world, I've placed an offering in a donation tray. In her version, I don't exist.

The waiter bows swiftly and turns away, walking down the pews of an old cathedral.

Back in the sunken city, I stare up at the stained-glass window until the dusk falls. ■

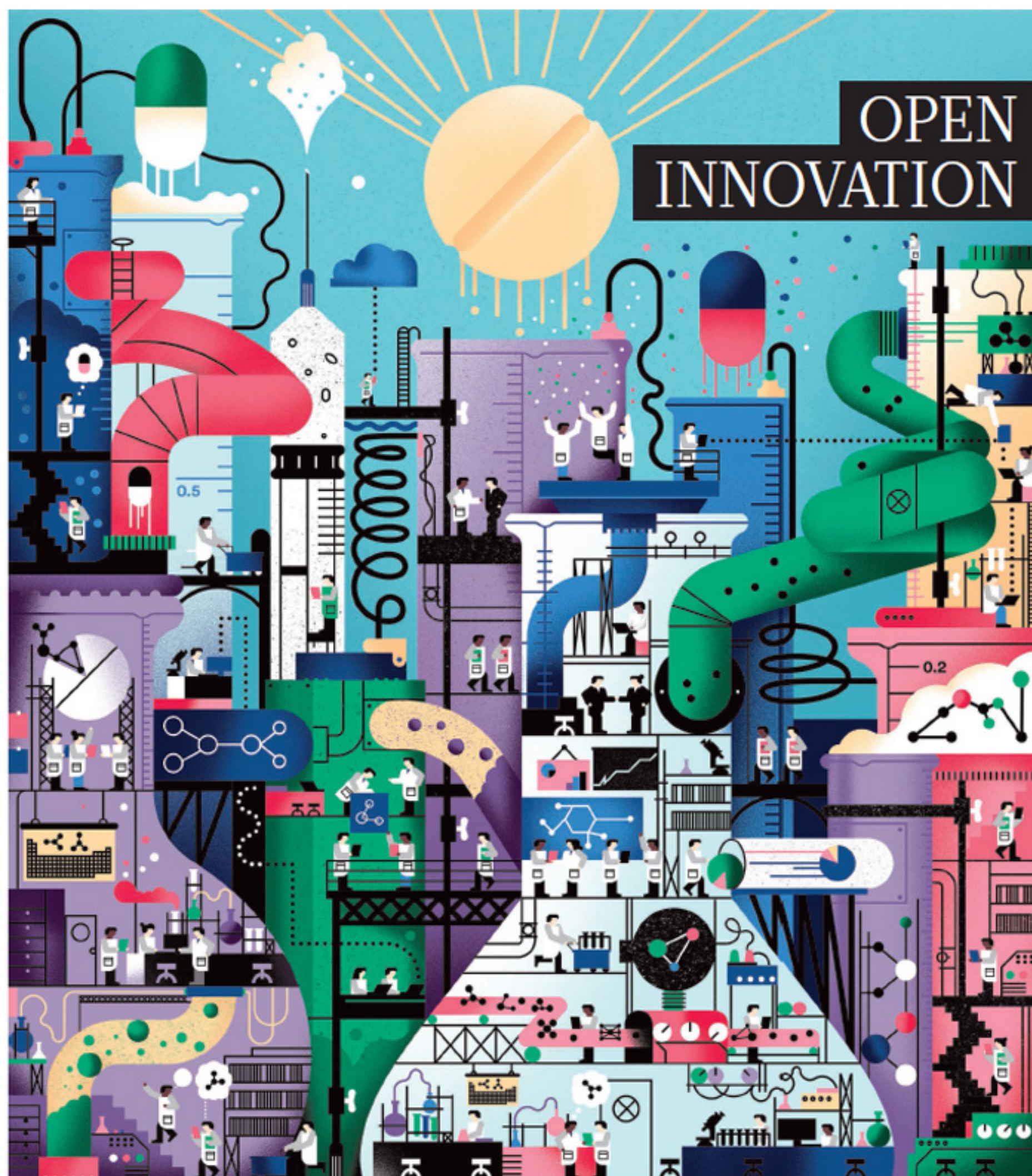
**Preston Grassmann** became a freelance writer after working as a regular reviewer for Locus Magazine. His most recent work has been published in AE: Canadian Science Fiction, Daily Science Fiction, Mythic Delirium and Slave Stories: Scenes From the Slave State.

ILLUSTRATION BY JACEY



12 May 2016  
Supplement to Nature  
Publishing Group journals

12 May 2016  
Supplement to Nature  
Publishing Group journals



Boehringer  
Ingelheim

# A transparent approach to biomedicine



# natureOUTLOOK

## OPEN INNOVATION

12 May 2016 / Vol 533 / Issue No 7602



Cover art: Bratislav Milenkovic

### Editorial

Herb Brody  
Michelle Grayson  
Richard Hodson  
Jenny Rooke

### Art & Design

Mohamed Ashour  
Andrea Duffy  
Wesley Fernandes

### Production

Matthew Carey  
Karl Smart  
Ian Pope

### Sponsorship

Stephen Brown  
Samantha Morley

### Marketing

Nicole Jackson

### Project Manager

Anastasia Panoutsou

### Art Director

Kelly Buckheit Krause

### Publisher

Richard Hughes

### Editorial director, partnership media

Stephen Pincock

### Chief Magazine Editor

Rosie Mestel

### Editor-in-Chief

Philip Campbell

Secrecy within the world of drug discovery and development is no longer as important as it once was. As development of therapies has become more difficult and costly, academics and industry competitors have begun to engage in greater collaboration and to embrace openness to accelerate research.

The road to a new drug is littered with expensive dead ends. Costly mistakes are often hidden from view and made multiple times in different laboratories (see page S54). To reduce this inefficiency, competitors are working together on basic research in pre-competitive partnerships (S56). Early research is being aided by the release of tools to explore potential drug targets. The freely available chemical probe JQ1, for instance, has sparked more innovation than it ever would have had it been kept locked away (S60). In the past decade, big pharma has come to accept that it must look for ideas beyond its walls if companies are to continue to innovate (S59). Hundreds of people, often without experience of the field, are entering competitions to solve complex biological problems (S62). And in an effort to screen the many compounds created each day, some companies have committed to examine the molecules without making claims over intellectual property (S65).

Research of neglected and tropical diseases is some of the most open. Open-source projects are publishing every step, and misstep, for all to see (S68). Despite these developments, some say that the life sciences are not as open as they should be (S70). One Canadian institute is preparing to test whether abiding by the principles of openness, including refusing to patent any of its discoveries, can work in a world where academics are expected to extract commercial value from their work (S71).

We are pleased to acknowledge the support of Boehringer Ingelheim in producing this Outlook. As always, *Nature* retains sole responsibility for all editorial content.

**Richard Hodson**  
*Supplements editor*

## CONTENTS

### S54 PROGRESS

#### A new chapter in innovation

How drug development is changing

### S56 COMPETITION

#### Unlikely partnerships

Rival pharma companies work together

### S59 Q&A

#### Change big pharma

Bernard Munos on how open innovation can bring drugs to market

### S60 CHEMICAL PROBES

#### A shared toolbox

JQ1's impact on innovation

### S62 CHALLENGES

#### Crowdsourced solutions

Global competitions to solve complex problems

### S65 COMPOUND SCREENING

#### Fresh hunting ground

Searching out drugs in the chemical haystack

### S68 TROPICAL DISEASE

#### A neglected cause

Lack of money prompts open research

### S70 PERSPECTIVE

#### Science is still too closed

Aled Edwards says society has work left to do

### S71 DATA SHARING

#### Access all areas

One institute's experiment in openness

## COLLECTION

### S73 Industry-academia collaborations for biomarkers

*Khusru Asadullah et al.*

### S75 Hit and lead criteria in drug discovery for infectious diseases of the developing world

*Kei Katsuno et al.*

### S83 A community-based approach to new antibiotic discovery

*Matthew A. Cooper*

### S85 Pioneering government-sponsored drug repositioning collaborations: progress and learning

*Donald E. Frail et al.*

### S94 Towards a hit for every target

*Steve Rees et al.*

*Nature Outlooks* are sponsored supplements that aim to stimulate interest and debate around a subject of interest to the sponsor, while satisfying the editorial values of *Nature* and our readers' expectations. The boundaries of sponsor involvement are clearly delineated in the *Nature Outlook* Editorial guidelines available at [go.nature.com/e4dwzw](http://go.nature.com/e4dwzw)

#### CITING THE OUTLOOK

Cite as a supplement to *Nature*, for example, *Nature* Vol. XXX, No. XXXX Suppl., Sxx–Sxx (2016).

#### VISIT THE OUTLOOK ONLINE

The *Nature Outlook Open Innovation* supplement can be found at <http://www.nature.com/nature/outlook/open-innovation>. It features all newly commissioned content as well as a selection of relevant previously published material.

All featured articles will be freely available for 6 months.

#### SUBSCRIPTIONS AND CUSTOMER SERVICES

Site licences ([www.nature.com/libraries/site\\_licences](http://www.nature.com/libraries/site_licences)): Americas, [institutions@natureny.com](mailto:institutions@natureny.com); Asia-Pacific, <http://nature.asia/jp-contact>; Australia/New Zealand, [nature@macmillan.com.au](mailto:nature@macmillan.com.au); Europe/ROW, [institutions@nature.com](mailto:institutions@nature.com); India, [npiindia@nature.com](mailto:npiindia@nature.com). Personal subscriptions: UK/Europe/ROW, [subscriptions@nature.com](mailto:subscriptions@nature.com); USA/Canada/Latin America, [subscriptions@us.nature.com](mailto:subscriptions@us.nature.com); Japan, <http://nature.asia/jp-contact>; China, <http://nature.asia/china-subscribe>; Korea, [www.natureasia.com/ko-kr/](http://www.natureasia.com/ko-kr/) subscribe

#### CUSTOMER SERVICES

[Feedback@nature.com](mailto:Feedback@nature.com)  
Copyright © 2016 Nature Publishing Group



# A NEW CHAPTER IN INNOVATION

A growing appreciation that cooperation and competition can coexist is transforming the life-sciences innovation landscape. Development was once shrouded in secrecy, but now organizations are coming together. By David Holmes; illustration by Mohamed Ashour.

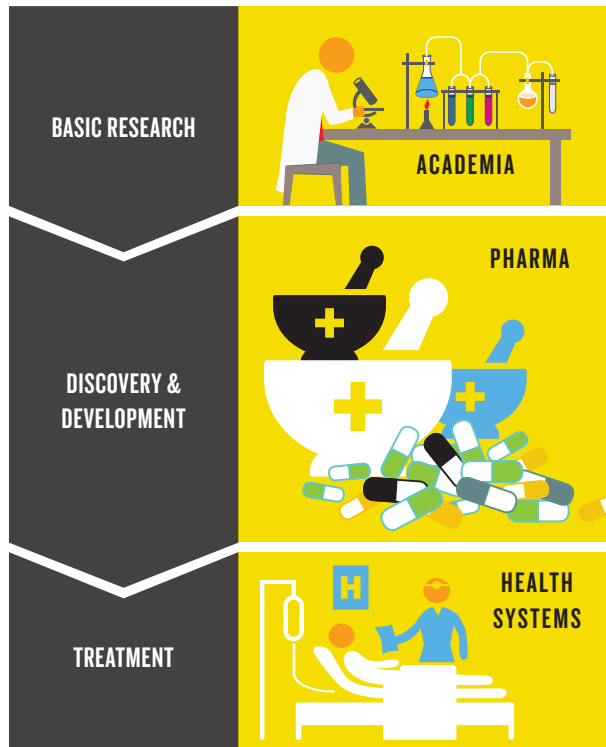
## STRAIGHT PATHS OF OLD

Historically, life-sciences partnerships were linear and closed in nature. The conception, development and delivery of therapies was handled separately — the involvement of one partner ending as that of another began.

This arrangement had its advantages. Risk and reward increased as the drug moved down the chain. But the lack of dialogue with the wider community could often result in costly duplications of effort.

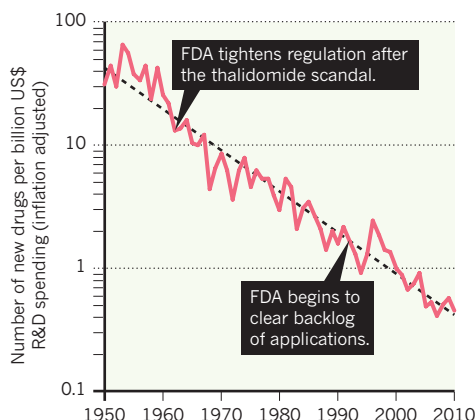
**US\$2.6 billion**

Average pre-tax cost per approved drug, including cost of failures<sup>1</sup>



## MOORE PROBLEMS

The number of drugs approved by the US Food and Drug Administration (FDA) per billion US dollars spent on research and development (R&D) has halved roughly every 9 years since 1950 — seemingly the opposite of Moore's law (Gordon Moore, co-founder of technology company Intel, observed that the number of transistors that can be placed, inexpensively, on an integrated circuit doubles every 18–24 months)<sup>2</sup>. In an effort to defy 'Eroom's law', organizations have been forced to search for ways to boost returns on investment.



## SQUEEZED MARGINS

Increasing R&D costs and competition from drugs that are no longer covered by patents have squeezed the margins of the conventional linear innovation model.

## INCREASING INFORMATION

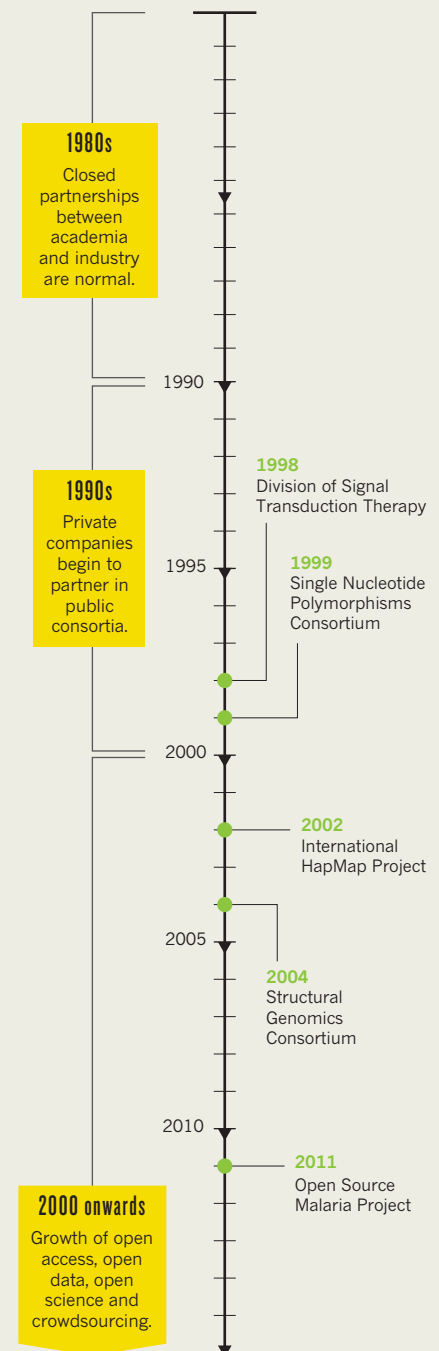
As researchers better understand the biological complexity of targeted diseases, and the technological complexity of drug discovery and the size of data sets increase, organizations are becoming more reliant on outside expertise.

## UNMET NEEDS

A conventional closed model of innovation is unsuitable for some medical needs, such as the development of antimalarials or Ebola vaccines.

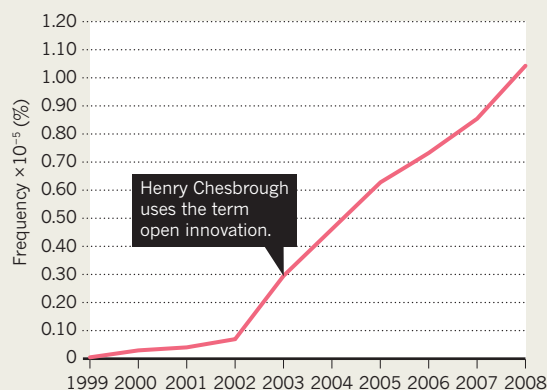
## A GROWING MOVEMENT

In response to the limitations of the closed model of innovation, a growing number of organizations have come together over the past two decades to share ideas, data and tools with each other, and often with the wider community.



## OPEN TO INTERPRETATION

According to Google's Ngram Viewer, which shows the relative frequency with which certain words and phrases are published, the phrase 'open innovation' gained prominence in the English language in 2003 — the year that business specialist Henry Chesbrough, now at the University of California, Berkeley, used it to describe the concept that companies should use external ideas as well as internal ones to advance their technology. But open innovation is just one of the terms that seeks to capture different, but overlapping, themes of openness in the broader innovation ecosystem.



### OPEN SOURCE

Pioneered by the software-development industry, open-source projects make the fruits of open collaboration free to everyone, with little or no direct profit for the collaborators.

### OPEN DATA

The Open Data Institute in London describes open data as data that anyone can access, use and share.

### OPEN ACCESS

Open access was a major shake up to the conventional publishing model. Literature is digital, online, free at the point of use and usually free to reproduce in some form.

### OPEN SCIENCE

Open science is the practice of science such that others can contribute. Lab notes are freely available, and the research and its underlying methods can be shared.

## DEGREES OF OPENNESS

### CASE STUDIES

Whether or not a collaboration qualifies as an example of open innovation can be subjective, but criteria **(pictured)**<sup>3</sup> devised by the Wellcome Trust in London can be used to categorize collaborations. It is generally agreed that the ideal open initiative would be a broad collaboration with public participation through all stages of innovative development, resulting in an output that is free to all.

Whom a collaboration is open to can vary, from an internal collaboration within an organization to a project fully open to anyone.

#### PARTICIPATION

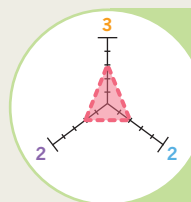
Open to the public 5

International collaborations with multiple partner types 4

Multiple partner types in a similar geography or limited number of international partners 3

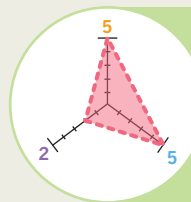
Limited number of partners (<10) within a similar geography 2

Internal collaboration within a single organization 1



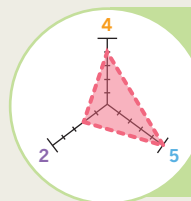
#### DIVISION OF SIGNAL TRANSDUCTION THERAPY

Industry researchers work with academics at the University of Dundee, UK, to understand the fundamentals of protein phosphorylation and ubiquitylation. The six participating pharmaceutical companies have access to unpublished results, technology and first rights to license intellectual property.



#### OPEN SOURCE MALARIA

The project takes promising public domain compounds and refines them for clinical trials. Anyone can contribute and all experimental data are shared in near-real time. E-mails are discouraged, with contributors instead urged to communicate on public discussion boards.



#### STRUCTURAL GENOMICS CONSORTIUM

The consortium is a public-private partnership that aims to catalyse drug-discovery efforts by determining the structure of potential drug targets. The structures are then released to the public, but no data are made available before then.

Narrow focus within a single R&D stage (usually very early stage) 1

Focus on early conceptualization (proof of mechanisms) 1

Pre-clinical or clinical studies (not both) 2

Pre-clinical and some clinical studies 3

Pre-clinical, clinical and commercial development 4

#### SCOPE

Openness can vary depending on the stage of the development process. Some projects are open until clinical studies begin, at which point access to data becomes tightly controlled. Open collaboration in other projects may persist through to final release.

Individual partner retains their intellectual property 1

Intellectual property is shared and retained within the partnership 1

Intellectual property generated before and as a result of the partnership is shared within the partnership only 2

Some of the resulting intellectual property is protected and made available to the public 3

A completely open approach is adopted in the partnership 4

#### ACCESS

For many, the right to access the fruits of a collaboration is central to whether it can be considered open. In some cases, individual partners retain intellectual-property rights, whereas in others results are made available to all.

Sources: 1. Tufts Center for the Study of Drug Development; 2. J. W. Scannell *et al.* *Nature Rev. Drug Discov.* **11**, 191–200 (2012); 3. R. Pigott *et al.* Shaping the Future of Open Innovation (Wellcome Trust, CASMI & Kinapse, 2014).



## COMPETITION

# Unlikely partnerships

*Drug discovery is time-consuming and full of blind alleys. Pharmaceutical rivals are cooperating in the early stages to accelerate and improve the efficiency of the process.*

BY NEIL SAVAGE

**T**he protein ubiquitin, as its name suggests, is found in almost all living tissue. It plays an important part in the death of old or damaged cells, by attaching to other proteins and labelling them for destruction. Failure to mark proteins in this way can lead to inflammation, cancer or neurological disorders such as Alzheimer's disease. If scientists can unravel the mysteries of this pervasive molecule, they may find new targets for drugs to treat these diseases.

Pharmaceutical companies are already selling three drugs that target processes involving ubiquitin as a way to treat the bone-marrow cancer multiple myeloma: bortezomib, approved by US regulators in

2003; carfilzomib, approved in 2012; and ixazomib, which got the nod last November. But because the system for attaching and detaching the protein has so many moving parts, including 2 activating enzymes, about 40 conjugating enzymes and some 600 ligases, there may be many more therapeutic targets still to be found. With so much to study, researchers at the University of Dundee, UK, and six pharmaceutical companies are collaborating to share their resources and findings, in the hope of gaining insights that will lead to new drugs. "I think this is going to become pretty big," says Philip Cohen, a biochemist at Dundee and one of the leaders of the collaboration, called the Division of Signal Transduction Therapy (DSTT). "The things we discover

here will be helpful to alleviate disease and also to generate a lot of money, not only for the companies, but for our own research."

The DSTT is a pre-competitive partnership — a type of collaboration in which pharmaceutical companies join together with one another, and often with academic researchers and the support of government funders, to tackle questions that they hope will lead to therapies. The idea is to share the cost of making early-stage discoveries, such as identifying biomarkers or disease pathways, that lay the groundwork for drug development. Armed with such basic knowledge, the companies can then identify specific molecules that might make drugs and study them in-house, developing proprietary therapies.

BRATISLAV MILENKOVIC



Beyond sharing costs, partnerships can also help with the sheer volume of biological information now collected. This will only continue to grow as DNA sequencing of individuals becomes more widely available. “The work is very large, and no single company or academic group can do it alone,” says Sylvain Cottens, who heads the Center for Proteomic Chemistry at the Novartis Institutes for Biomedical Research in Basel, Switzerland. If academic and industry researchers can pool their resources and share skills, they may be able to improve efficiency and speed up the creation of therapies for a wide variety of diseases.

### THE COST OF FAILURE

Most drug candidates go nowhere. In 2004, the US Food and Drug Administration (FDA) estimated that only 8% of the compounds that enter phase I trials — many of which have been in development for more than a decade — actually make it to market. In 2015, the industry group Pharmaceutical Research and Manufacturers of America put that figure at less than 12%. The average cost of developing a drug in the first decade of the twenty-first century was US\$2.6 billion — up from an average of \$1 billion in the 1990s (see ‘Under pressure’).

Pre-competitive partnerships could be a way to dramatically improve the efficiency of drug development. For starters, they could reduce the large amount of duplication. Companies conduct their research in secret and tend not to publish the results of failed studies, meaning that other groups are likely to follow the same fruitless lines of inquiry. “If ten companies are working on Alzheimer’s disease on exactly the same target and it’s failed, that’s ten times the investment that is down the tubes,” says molecular biologist Pierre Meulien, who heads the Innovative Medicines Institute (IMI) in Brussels, a partnership between the European Commission and the European Federation of Pharmaceutical Industries and Associations. Because of the secrecy, it’s difficult to come up with specific figures for duplication. But in 2009, it was estimated that 85% of research is wasted (amounting to \$170 billion worldwide each year), at least some of which is because of failed or redundant studies (P. Glasziou & I. Chalmers *Lancet* 374, 86–89; 2009).

Greater openness could reduce redundancy and save money, as well as spare patients from enrolling in trials that are doomed to fail. But as Cohen sees it, the real promise of pre-competitive partnerships is improving our understanding of the biological mechanisms that underlie a

**“The patent system has not been designed with open collaboration in mind.”**



Protein structures solved by the Structural Genomics Consortium are made publicly available.

particular process. The DSTT’s focus on deubiquitinating enzymes, which modify the effect of ubiquitin, is already helping to speed up the discovery of candidate drugs. One company, Dundee-based Ubiquigent, has already been formed to provide drug-development companies with assays and reagents developed by researchers at the University of Dundee. Cohen hopes that deubiquitinating enzymes will follow the path of kinase inhibitors, which the DSTT also studies. Once the first kinase inhibitor, imatinib, was approved by the FDA in 2001 to treat chronic myeloid leukaemia, researchers began devoting more resources to them. Since then, more than 25 drugs that target kinases (which help to control the function of certain proteins) have been approved.

The DSTT, which was formed in 1998, is made up of AstraZeneca, Boehringer Ingelheim, GlaxoSmithKline, Janssen Pharmaceutica, Merck KGaA, Pfizer and 20 academic research teams, and according to Cohen is probably the longest running collaboration of its kind. Under the terms of the agreement, all unpublished results are shared between the collaborators, along with reagents, technology and technical know-how. Faculty and students must sign confidentiality agreements regarding the companies’ intellectual property, although they can still publish papers based on the collaboration’s research. The first drafts of articles are shared on a private website. Any member that wants a head start on development and patenting before the information reaches the public has 45 days to request a 9-month publication delay. Cohen says that the number of papers delayed is low — perhaps around 10 out of the past 400 — and that in practice, the delay is not as long as it seems because researchers tend to share drafts at earlier stages than they

would submit them to journals. “We actually think that the threat of this delay has caused us to publish more effectively and efficiently,” he says.

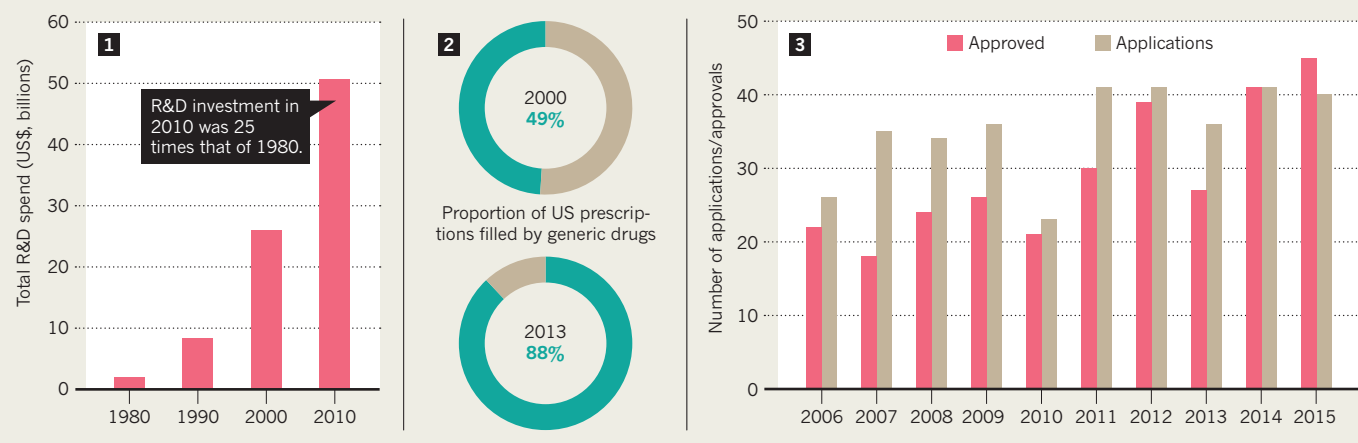
By contrast, the public-private partnership the Structural Genomics Consortium (SGC) has no members-only viewing period. Its policy is to release data to its members and the rest of the world simultaneously, with no restrictions on use. The SGC also promises never to patent anything. “That openness does lead to faster science,” says Aled Edwards, a protein biochemist at the University of Toronto, Canada, and founder of the SGC. The SGC determines a protein’s structure and publishes the information in the international repository the Protein Data Bank. Under the policies of many scientific journals, anyone who describes a protein structure must deposit their data in the bank to make it available to all researchers. In the past 12 years, the SGC has deposited more than 1,500 descriptions of protein structures, from both humans and parasites, into the data bank. It also develops antibodies and chemical probes — small molecules that test how a potential drug interacts with biological targets (see page S60).

SGC members all have different strengths. Academic researchers, Edwards says, are good at making basic discoveries, but have no incentive to take them beyond published papers. Participating pharmaceutical companies, however, are much more focused on creating marketable therapies. They are very good at high-throughput screening of drug candidates, but don’t spend much time on the most basic science. “We do get knowledge from the academic groups, but we also provide knowledge about how to develop assays,” says Cottens — Novartis is an SGC partner.



## UNDER PRESSURE

The pharmaceutical industry estimates that it is now more expensive to bring a drug to market than ever. Research and development (R&D) investment by members of the trade group the Pharmaceutical Research and Manufacturers of America was more than US\$50 billion in 2010 **1**. Pressure from generics (identical or equivalent versions of drugs for which patent protection has expired) that are typically substantially cheaper than branded options is also increasing **2**. But the number of drugs coming to market each year is rising — 2014 and 2015 saw the highest number of US Food and Drug Administration approvals in the past ten years **3**.



SOURCE: LEFT AND MIDDLE: PHRMA; RIGHT: FDA

Unlike the DSTT and the SGC, the IMI has no general rules about intellectual property. Instead, Meulien says, details of what can be shared and what stays proprietary are agreed in advance between collaborators on a given project. “We have a whole spectrum of types of arrangements.”

Between 2014 and 2024, the IMI will receive €1.6 billion (\$1.82 billion) from the European Commission and €1.4 billion from European pharmaceutical companies. This will fund projects that focus on neurological conditions, diabetes, cancer, tuberculosis, obesity, vaccine safety, the use of stem cells in drug discovery and antimicrobial resistance, among others. One IMI effort, the European Lead Factory, provides small- and medium-sized companies, as well as academics, with free access to half a million chemical compounds, which they can use to screen potential drug targets.

### NOT HITTING THE PRICE POINT

Although a promising approach to point the way to therapies, partnerships that focus on fundamental science may have little impact on the overall cost of research. This is because it is during clinical trials, rather than early research, that most drug candidates fail. The SGC, for instance, is “at a fairly inexpensive part of the pharmaceutical discovery process,” Edwards says.

So, in 2011, he and SGC chief scientist Chas Bountra, a translational medicine specialist at the University of Oxford, UK, joined with Sage Bionetworks, a non-profit biomedical research organization in Seattle, Washington, to form a partnership they called Archipelago to Proof of Clinical Mechanism (Arch2POCM). The idea was to extend pre-competitive cooperation on a few drug targets into phase II clinical trials, after which the risk of failure drops substantially.

But the vision proved too ambitious, says chemist Thea Norman, Sage’s director of strategic development. She says that the pharmaceutical companies worried that it might prove too difficult to base intellectual property on compounds and information that would be in the public domain. “The idea was a new one and one that maybe at first glance for a pharmaceutical company takes a little explaining,” she says. “We had at least two pharmaceutical companies that were ready to sign up, but we felt we needed a little more critical mass than that.” To get enough funding for what they had in mind, she says, they needed three to five companies on board.

When the first approach turned out to be more than the industry was willing to sign up to, Arch2POCM’s founders launched a smaller-scale project, but one that still went beyond previous collaborations. The group began a 3-year UK effort in late 2013 with the Institute of Cancer Research and Newcastle University, with funding from Cancer Research UK and the Avon Foundation for Women, but with no pharmaceutical companies involved. The scaled-back programme aims to find a candidate compound that works on the enzyme KDM4B, which is implicated in people with breast cancer, but won’t take the compound all the way through phase II trials. Norman says that the hope is that, whatever the scientific outcome, the project will demonstrate that cooperation can benefit drug development beyond the earliest stages of discovery.

Another way to maintain the openness between pharmaceutical companies further into the drug-development process

*“We hope to really transform the ecosystem for how these things are done in Europe.”*

may be to change financial incentives. Liza Vertinsky, who focuses on intellectual property at Emory University in Atlanta, Georgia, says that current patent law encourages companies to jealously guard their intellectual property, because if they lose patent protection they could lose out on the profits that come with a successful drug. “The patent system has not been designed with open collaboration in mind, so the mechanism of how you would share intellectual property is not built into the system,” she says. An alternative would be for lawmakers and courts to develop a concept of fair use, analogous to laws that allow people to quote a passage from a novel or sample a snippet of a song while not violating copyright, for example. That way, she argues, companies could share some portion of their research findings without giving up all claims to their intellectual property. Vertinsky intends to look more closely at pre-competitive partnerships in the coming year to better understand how changes in the law might affect the way they work.

Even if the cooperation between pharmaceutical companies cannot be expanded further, pre-competitive partnerships are still having a positive effect on drug discovery, participants say. Although Cottens won’t go into specifics about propriety work, he says that the collaboration “has clearly accelerated” the projects that Novartis is working on. Meulien says that these efforts are already helping to translate academic knowledge into practical applications. “We hope to really transform the ecosystem for how these things are done in Europe,” he says. “We do things that no one company or university could do alone.” ■

Neil Savage is a freelance writer based in Lowell, Massachusetts.



**So, is the process now less dependent on the financial power of the big companies?**

The drug-discovery process has become a lot more open. You can be a one-person pharmaceutical company today. You have access to many of the same databases and resources as your colleagues in the pharmaceutical industry. You can also use crowdsourcing platforms to get the input you need at a very attractive price. I know one company that went from creation through to phase II clinical trials with only 6 people and about US\$70 million in cumulative funding. That is a puny amount compared with what it used to cost.

**How will the move towards patient-gathered data disrupt the industry?**

Biosensing technologies allow patients to be monitored at home, generating vast amounts of high-quality data at very low cost. The impact of these technologies on clinical research will be huge, and the speed at which it is happening is astonishing. These technologies will change the industry. They will reduce the cost of research and erase much of the advantage that scale used to confer on big companies.

**What happens when patients take a more active role in the control of their clinical data?**

This will be game-changing. Whoever controls data collection and access will ultimately control innovation. Patients realize that they can play a central part there, and are organizing themselves accordingly. This is quite a threat for industry, because patients have different values to pharmaceutical companies. Patients are pushing their values on the clinical research enterprise, and they are values of openness, speed and convenience.

**And the value of affordability?**

Yes, the cost of innovation is still a big problem. Of the pharmaceutical needs in the United States, 90% are met by generic drugs costing \$70 billion; the remaining 10% cost \$350 billion. That should be concerning to pharmaceutical leaders. Their new drugs often treat only thousands of patients, whereas the old blockbusters treated millions. And how can you grow a \$350-billion revenue base if you have lost millions of customers?

**Do drug companies fully understand this crisis?**

Most do not produce enough innovation to grow. In fact, half of them are shrinking. They try to mitigate this by escalating prices, which is dangerous. I think industry is misjudging the anger that its practices are creating. The risk of a backlash is real. Policymakers could quickly make changes that may be much less palatable to industry than what could have been achieved by self-policing. ■

**INTERVIEW BY ERIC BENDER**

This interview has been edited for length and clarity.

## Q&A Bernard Munos

# Change big pharma

*In 2006, pharmaceutical innovation consultant Bernard Munos helped to launch a lively public discussion about how open innovation can bring novel drugs to market with his paper 'Can open-source R&D reinvigorate drug research? He tells Nature how things have changed since then.*

**How has the research landscape changed over the past decade?**

Until ten years ago, innovation was controlled by the big pharmaceutical firms, because they were the only organizations able to afford it. Today, there are all kinds of new players and research models that have spread innovation to all parts of the ecosystem.

Big pharma used to be inward-looking about their research pipelines; overwhelmingly, their ideas came from inside the company. But the number of drugs approved was too low to secure the future of the industry. Companies realized that they needed more, better and cheaper innovation. Now, there is a lot of openness. Some firms have adjusted their research model to effectively tap the global brain. The most successful are willing to contemplate totally new science.

**What differentiates the companies that do so?**

It comes down to leadership. Take Paul Stoffels, chief scientific officer of pharmaceutical company Johnson & Johnson (J&J). He championed the idea of Janssen Labs, an outreach organization that is run alongside J&J's research. Janssen Labs reaches out to those who do breakthrough science, and lets them know that they can avail themselves of the

scientific, and potentially financial, support that J&J can provide. It hosts these partners in innovation incubators, which nurture dozens of companies with the potential to come up with game-changing therapies. The drug candidates born of these projects have not yet percolated through the J&J pipeline, but when they do, the company's drug output could be multiplied several-fold.

**How else is drug innovation changing?**

You've got all kinds of groups doing viable research now. Universities are tired of seeing their great ideas languishing on the shelf because they fall outside the comfort zone of big pharma. They want to give those ideas a fair chance, so they are engaging in translational research themselves.

Patient advocacy organizations are taking things into their own hands and launching credible drug research and development (R&D) programmes that run on a shoestring.

Venture philanthropists, who for personal reasons may have an interest in certain diseases, are willing to part with large amounts of money to fund initiatives to help with those diseases. All of this not only enlarges the global pipeline, but also changes the economics of drug R&D.





James Bradner (left) who worked at the Dana-Farber Cancer Institute in Boston, Massachusetts, with Jun Qi who synthesized the JQ1 compound.

## CHEMICAL PROBES

# A shared toolbox

*In a pioneering move, the compound JQ1 was released to the community for free. The impact that this has had on research and development is slowly coming into focus.*

BY ANDREW R. SCOTT

In 2010, a young New England firefighter lay dying from a rare and aggressive form of cancer. The disease, NUT midline carcinoma, had taken hold in his left lung and spread widely. A chest tube, inserted to offer some comfort, continually drained fluid containing cancer cells. When doctors explained the research potential of these cells, the man readily agreed to donate them to scientists at the Brigham and Women's Hospital and at the Dana-Farber Cancer Institute, both in Boston, Massachusetts. This gift was one of the key steps in a pioneering development in open innovation that now serves as a model for drug researchers worldwide.

James Bradner's group at Dana-Farber had a molecule that they hoped might make these

cells "forget they were cancer", as Bradner puts it. Sure enough, on exposure to this molecule the cells seemed to return to normal. And when the cancer cells were grown in mice — a model of the firefighter's disease — and the animals were treated with the molecule, the tumours began to shrink (P. Filippakopoulos *et al.* *Nature* **468**, 1067–1073; 2010).

Sadly, this result could not save the firefighter, but Bradner's decision to broadly release the compound through an open-source model of drug discovery has spawned a raft of patents and clinical trials of a new class of drugs: bromodomain inhibitors.

The molecule that affected the firefighter's cancer cells so dramatically was JQ1, named after the chemist who made the original compound, Jun Qi. JQ1 was not intended to be a drug; limitations in its solubility and half-life

mean that it probably never will be. Rather, JQ1 is a chemical probe — a small molecule that interacts with a target protein to allow the activity of that protein to be investigated in the cell.

JQ1 binds to a pocket-shaped region found in many proteins, called a bromodomain. This structure detects a small molecular signal attached to the histone proteins that parcel up DNA into chromatin. When the signal, which Bradner describes as "a molecular post-it note", is recognized by the bromodomain, it can stimulate the protein to begin a series of interactions that results in the activation of genes involved in sustaining cell growth and activity. But in some diseases, including cancer, cells interpret this signal as an instruction to grow and divide out of control. The development of compounds to stop

this errant bromodomain signalling could be a focus for therapy, and, as Bradner saw it, JQ1 had the potential to kick-start the process.

### OUT IN THE OPEN

When JQ1's effect on the firefighter's cancer cells was first discovered, the conventional approach would have been to keep everything secret, explains Bradner, who took up the role of president of Novartis Institutes for BioMedical Research, based in Cambridge, Massachusetts, in March. Secrecy was the standard approach — all the details were kept hidden until either the prototype drug had been turned into an active pharmaceutical compound or efforts had come to a standstill. "But we did just the opposite," says Bradner. In what he describes as a "social experiment", he and his laboratory released all the information they had on JQ1. Not only that, but they also promised to supply unlimited quantities of JQ1 to any researcher who wanted it, for free and without restriction on use.

Bradner believes that this approach has greatly accelerated drug development in the field of bromodomain inhibition. His lab has sent samples of JQ1 to more than 400 laboratories worldwide, both in academia and industry, and there has been a clear increase in research activity around the bromodomain proteins it targets, leading to more than 100 filed patents (see 'Driving innovation'). These patents are not for JQ1 itself, but for other molecules that target bromodomains — the development of many of these was guided by the use of JQ1 as a research tool. Some of the structures of these drug candidates are similar to JQ1, but others are completely unrelated. Bradner is convinced that, had he not adopted the open approach, there might be only one, or at most two, bromodomain inhibitors in clinical trials, including his own, TEN-010. Instead, he knows of eight agents under development for the treatment of cancer alone.

"Normally the process of obtaining agreement to use these valuable reagents can be long and complex," says physician-scientist Ross Levine at the Memorial Sloan Kettering Cancer Center in New York. Being able to obtain probes such as JQ1 through open-access agreements has, he says, greatly assisted his research.

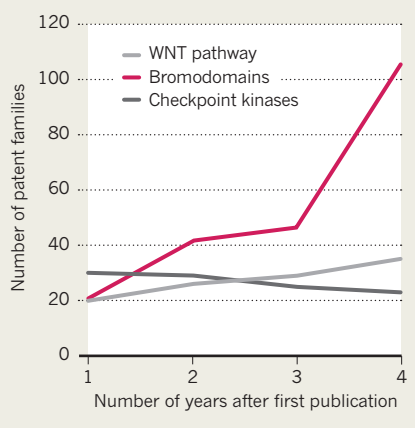
Researchers who attempted to assess the impact of JQ1 came to a similar verdict (Z. Arshad *et al. Expert Opin. Drug Discov.* 11, 321–332; 2016). They found that the open release of JQ1 increased innovation in the broader field of bromodomain inhibition — evidence, they said, that an open-access approach can improve the efficiency and lower the cost of both drug discovery and commercialization.

### THE WIDER VIEW

Clinical trials are now under way on the use of bromodomain inhibitors to treat some of the most common cancers, including lymphomas, leukaemias, multiple myelomas and solid

### DRIVING INNOVATION

More patents have been filed related to bromodomains, the target for the JQ1 chemical probe, than for similar drug-discovery targets for which the probes were not made openly available.



tumours (such as cancers of the pancreas and prostate). Development of all these molecules has benefited from data gleaned from the availability of the molecular probe.

And cancer is not the only target. The open release of JQ1 is also proving to be a boon for research into various neurodegenerative disorders such as Alzheimer's disease, as well as inflammation and viral infection.

JQ1 is just one molecule, however. How much of a trendsetter it will prove to be remains to be seen. Structural biologist Wen Hwa Lee at the University of Oxford, UK, says that although the progress made with bromodomains is a sign of the value of open access, the wider impact of open innovation remains limited. "It is such a new concept," he says. "Having been told all their careers that they have to protect their data, many scientists can find it hard to get their heads around the idea of openly sharing."

Still, developers of other chemical probes are aiming to replicate the JQ1 success story. As part of the drive to increase the ease of access to many more probes, an alliance of researchers, with the endorsement of three life-sciences research organizations — the Structural Genomics Consortium (to which Lee belongs), the Institute of Cancer Research in London and the Broad Institute of Harvard and Massachusetts Institute of Technology in Cambridge, Massachusetts — have been leading an effort funded by the UK's Wellcome Trust to collate data on chemical probes and allow researchers to obtain supplies of these molecules. In late 2015, the group launched chemicalprobes.org, an online portal to freely share probe data. Anyone can submit a probe for publication on the site, although each submission is subjected to peer review — designed to address past problems with poor-quality probes that generated misleading research results. As of March 2016, the portal hosted more than 100 probes targeted against 10 key

families of proteins, including bromodomains. Many of these probes are already being used to guide researchers towards new drugs for clinical trials.

### COMMERCIAL INTEREST

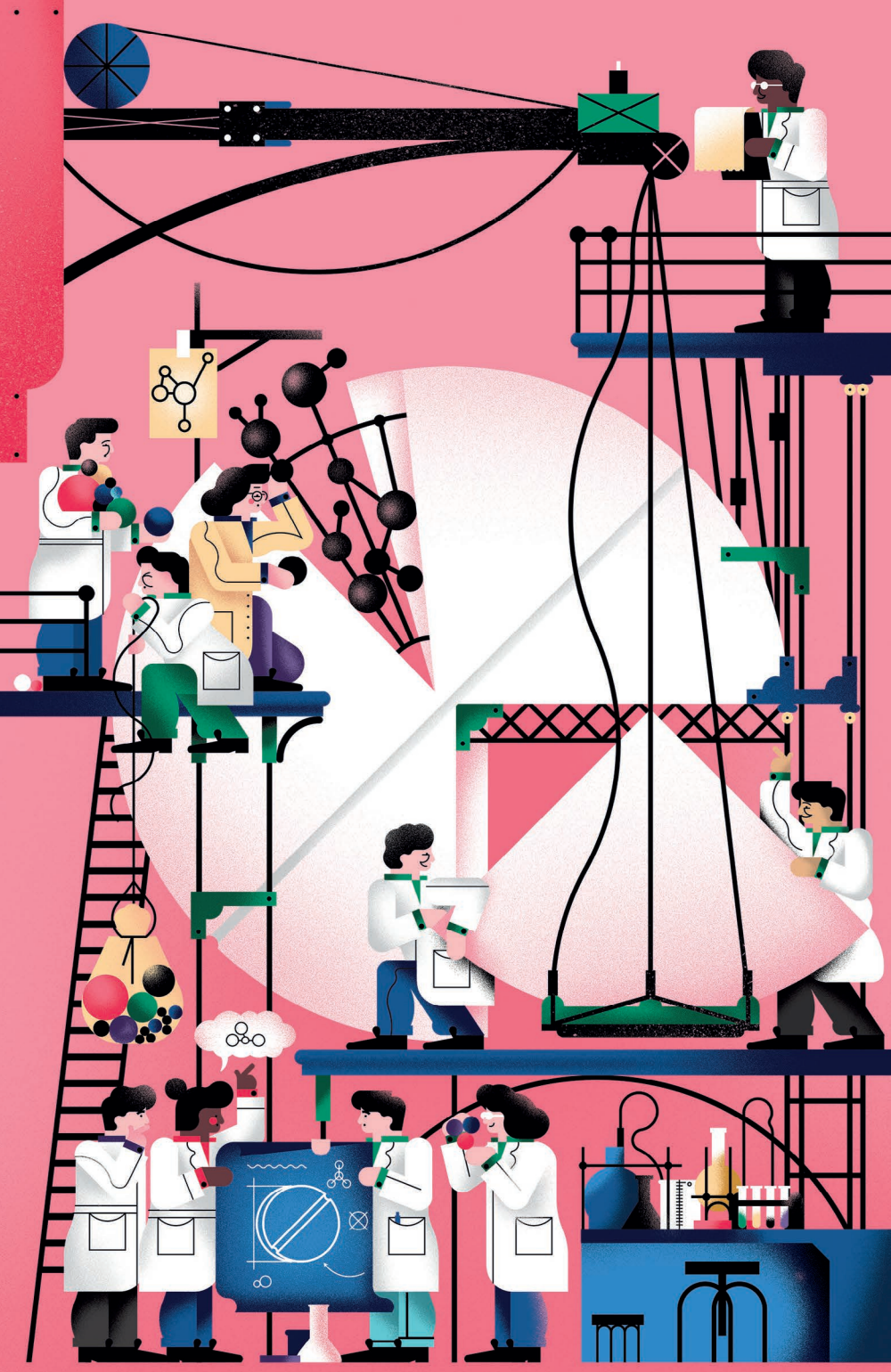
Sharing chemical probes doesn't just appeal to academics — the practice is catching the attention of drug companies too. Mark Bunnage, a medicinal chemist at biopharmaceutical company Pfizer, believes that JQ1 has had "a profound impact" on basic research that has allowed the search for new drugs. As well as making use of the probe itself, Pfizer has also begun to release more information about its own molecular tools. Many of the company's chemical probes are available through vendors such as life-sciences company Sigma Aldrich. Pfizer also collaborates with the Structural Genomics Consortium to create new probes, a number of which have been included on the chemical probes portal. "This is a big change of mindset," says Bunnage. "Open-innovation approaches are definitely helping companies and academics to work together to advance the discovery and development of medicines." He hopes that the increase in collaboration in the earliest stages may reduce the number of prospective drugs that fail during phase II clinical trials.

The key to collaboration between academia and life-sciences companies is the sharing of chemical probes so that a protein or other biomolecule can be identified as a drug target early. This accelerates the basic science and allows market dynamics to kick in to incentivize the development of specific drugs. "Everyone can eventually try to develop the best drug to interact with each target," says Bunnage. Lee agrees that for open-research platforms such as the chemical probes portal to be part of a successful economic model, there comes a point at which barriers of confidentiality must come down. At some stage drug companies, driven by their requirement to make profits, must be free to develop their candidate drugs confidentially. Negotiating the transition from open to closed innovation may be a difficult process with high stakes. "Where the pre-competitive line lies must be explored by the partners involved and may also depend on the disease area and its needs," says Lee (see page S56).

The path that Bradner's career has taken since his decision to release JQ1 into the wild demonstrates that openness and commercial interests are not mutually exclusive. Tensha Therapeutics, a start-up he founded in 2011 to develop drug-like bromodomain inhibitors, was acquired by pharmaceutical company Roche in early 2016 in a deal worth around US\$500 million. Clearly, acting in the interests of open science need not preclude commercial success. ■

**Andrew R. Scott** is a science writer based in Perth, UK.





BRATISLAV MILENKOVIC

Hours before the close of Kaggle's competition to find out why almost one-third of women in the United States are not screened for cervical cancer, the leading team has submitted the 115th iteration of its model. Forty groups around the world are competing to win US\$100,000 in a challenge sponsored by biotechnology company Genentech.

The models are based on analyses of a 150 gigabyte database of de-identified patient data, says computational biologist Wendy Kan, who set up the challenge and works at Kaggle in San Francisco, California, a company that runs predictive modelling and analytics competitions that allow data scientists to compete to solve complex problems. In addition to finding solutions, contestants are asked to explain their reasoning. "It's very important for us to tell a story," Kan says. Later, on a Kaggle forum, a member of the winning team presents two of the group's hypotheses: multiple chronic diseases and mental-health issues are major factors in why some women skip screening.

Another Kaggle challenge, which began in December, asked participants to transform the diagnosis of heart disease by coming up with an algorithm to examine cardiac magnetic resonance imaging (MRI) scans to see how well the heart is pumping blood — "A very difficult problem," Kan says. Entrants used a cardiac MRI data set provided by the US National Heart, Lung and Blood Institute, and 192 teams were in the running for the \$200,000 prize when the competition closed. The victors were two quantitative analysts who have worked with hedge funds, but had no experience in cardiology.

So far, more than 450,000 data scientists have tried their hand at Kaggle's predictive-modelling puzzles, says economist Anthony Goldbloom, founder and chief executive of the organization. The problems — many pertaining to health, but others in fields that range from criminology to search technology — are set up so that the background of entrants doesn't matter, he says. As long as they have suitable modelling skills, no particular experience or qualifications are needed.

"They are all smart, highly motivated and incredibly capable," adds Goldbloom. "The winning margin is usually very small; often the difference between first and second isn't even statistically significant."

Kaggle is one of a number of organizations running open global challenges in life sciences to address knotty problems in basic biology, clinical research or health care. The approach is steadily gaining backers in academic laboratories and classrooms, drug companies and government agencies as a way to bring well-defined, but thorny problems to the attention of brilliant minds around the world.

The design of the competitions varies from challenge to challenge and host to host. Some ask for modelling algorithms, others for ideas, and still more for prototype medical solutions.

## CHALLENGES

# Crowdsourced solutions

*Open competitions bring new minds, skills and collaborations to problems in biomedical research.*

BY ERIC BENDER



Prizes are often offered, although participants usually insist that money is not the main motivation. Some of the winning solutions, especially those sponsored by industry, remain secret, but others are made openly available and a few have already resulted in advances in clinical research.

### CLOCKWORK ORIGINS

Competitions in science and engineering have a long history. In 1714, the Longitude Act saw the UK government offer a reward of £20,000 (well over £2,000,000 (\$2,865,400) in today's money) for a solution to the problem of calculating longitude at sea. Not just one, but two answers emerged: the marine chronometer, developed by clockmaker John Harrison, which kept time at sea well enough for navigators to calculate longitude effectively; and a method for devising longitude from the motion of the Moon borne of a combined effort by scientists, including mathematician John Hadley and astronomer Tobias Mayer.

But it took the advent of the Internet for crowdsourced medical contests to really take off, notably with the Critical Assessment of Protein Structure Prediction (CASP) experiments, which have seen research groups test their methods for predicting 3D protein structures against those of their peers since 1994.

The competitions gained more industry backing as pharmaceutical companies began to struggle with their pipelines. The crowdsourcing firm InnoCentive, for example, formed in 2001, a time when “the pharmaceutical industry needed to rethink its business model”, recalls Alph Bingham, co-founder of the company based in Waltham, Massachusetts, and then a vice president at pharmaceutical giant Eli Lilly. “The Internet let you access minds on a scale and a scope that had never been possible before.”

**“Prizes can really bring a new population of researchers into the field.”**

Spun out from Eli Lilly, InnoCentive has held more than 2,000 open challenges and attracted more than 375,000 ‘solvers’. The continual string of challenges can be tightly focused and relatively small, such as a \$30,000 challenge to find a minimally invasive skin-biopsy method to measure gene expression, or attempt to tackle larger problems, such as a major \$500,000 challenge sponsored by the US National Institutes of Health (NIH) to look for robust methods to examine individual cells. Proposals such as these are inherently risky and might not survive the conventional NIH grant process.

Indeed, challenges seem to hold a number of advantages over conventional research practices. One of the leading crowdsourcing initiatives is the Dialogue for Reverse Engineering Assessments and Methods (DREAM) Challenges programme, which sees groups compete in open competitions to solve complex



Scientists from around the world competed to win a BioMed X fellowship in Heidelberg, Germany.

modelling problems in systems biology, says Gustavo Stolovitzky, co-founder of the project and computational biologist at IBM's Thomas J. Watson Research Center in Yorktown Heights, New York.

When dozens of teams around the world take on a DREAM project, they often accomplish in months what would take a single research group years, “since you can multiply the number of people working on the problem by 50 or 100,” says Stolovitzky. Many challenges also bring in researchers from other fields, who may approach problems in ways that those closely acquainted with them would not.

Just as crucially, challenges jump-start collaborative communities. For instance, the ICGC-TCGA DREAM Somatic Mutation Calling Meta-pipeline Challenge is a collaboration between DREAM, the International Cancer Genome Consortium, The Cancer Genome Atlas and biomedical research organization Sage Bionetworks in Seattle, Washington. Its aim is to improve standard methods for identifying cancer-associated mutations and rearrangements in whole-genome sequencing. In the process, they are building an ongoing community in which researchers can find the best and latest algorithms, rather than having to go to scientific journals.

Crowdsourced tournaments can also open up access to data — either those aggregated specifically for the purpose, such as Kaggle's cervical-cancer and cardiac MRI databases, or data sets that would otherwise lie dormant. “There are too many data silos in which researchers hoard their data, sometimes for years,” Stolovitzky says. “Ultimately, everybody should be able to look at that data with information about how the data was gathered, allowing collaboration and data

sharing in a positive and meaningful way.”

In addition, contests can lower the legal barriers that plague collaborations between institutions or companies, says Bingham. “They offer ways to engage all these different people without having to precede that whole process with 200 days of legal briefs being exchanged between institutions,” he says.

For these contests to achieve these positive impacts, however, they have to be well managed. Crowdsourcing is of little help in areas in which research is at such an early stage that the organizers can't ask the right questions. For any challenge to work, the problem must be well-defined and able to be judged fairly, says systems biologist Stephen Friend, co-founder and director of Sage Bionetworks. It's also important for an impartial expert in the field to act as a convener and nurture the emerging community, he says.

Non-profit foundations — increasingly important providers of research funding — are also making use of crowdsourcing. Often these focus on diseases that drug companies rarely target (see page S68). One example is Prize4Life in Berkeley, California, founded in 2006 when Harvard business school graduate Avichai Kremer was diagnosed with amyotrophic lateral sclerosis (ALS; also known as motor neuron disease), and best known for its \$1-million contests.

“Prizes can really bring a new population of researchers into the field,” says neuroscientist Neta Zach, chief scientific officer at Prize4Life. “And a lot of them continue to work on ALS.” Prize4Life's first major challenge addressed the lack of useful biomarkers for ALS progression. “We expected that the tool would be based on measurements from blood or cerebral spinal fluid,” Zach says.





Participants at a Massachusetts Institute of Technology Grand Hack discuss health-care challenges.

Instead, the winning tool in 2011 was a more creative solution: a pain-free non-invasive medical device that measures the flow of electrical current through muscle tissue. The winnings helped to build the San Francisco start-up Skulpt, which is testing such devices in ALS trials (as well as offering them to consumers as fitness tools).

The foundation also partnered with DREAM and InnoCentive in a \$50,000 challenge to predict the progression of ALS. When the predictions of the winning algorithm were compared with those made by ALS clinicians in the assessment of 14 people with ALS (R. Kuffner *et al. Nature Biotechnol.* 33, 51–57; 2015), “the algorithm outperformed each and every one of the clinicians on each and every one of the patients”, Zach says. The model is now used to make ALS clinical trials more efficient and their results clearer — a better understanding of ALS makes it easier to assess the benefits of treatment.

DREAM was launched in 2006 by Stolovitzky and systems biologist Andrea Califano at Columbia University in New York City to improve the state of the art in systems-biology modelling. As well as solving problems, DREAM challenges validate the solutions.

Sometimes when data-science groups tackle a difficult problem, they can convince themselves that they have produced a good solution, rather than actually solving it well. Stolovitzky calls this the “self-assessment trap”, which can lead to mistakes such as overfitting models to one set of data. But if 50 DREAM teams are involved, “we can see if we can really find a clear signal in the data”, he says.

In 2012, DREAM joined forces with Sage Bionetworks, which had created Synapse, a pioneering open-computing platform for data analysis and sharing. The first joint challenge generated models to classify the aggressiveness of breast cancer. The models clearly performed better than today’s commercial tests, says Friend. “More importantly, the challenge showed that people who had not generated the

data were able to get deep insights,” he says. “And the electrical engineer who won had very little chemical background.”

### RISE TO THE CHALLENGE

Competitions are beginning to exploit the opportunities provided by data contributed directly by patients. Sage, for example, created mPower, an app that uses iPhone sensors to measure symptoms of Parkinson’s disease progression such as dexterity or gait. And Sage has partnered with other groups, such as Oregon Health and Science University in Portland and Harvard University in Cambridge, Massachusetts, to create numerous such apps, which

**“At the end of the day, cash is often a scorecard, not a paycheck.”**

can very quickly provide high-quality data. “We have over 200,000 people who have said, I want to share my data with qualified users,” Friend says.

In November 2015, a DREAM hackathon drew participants for two evenings of pizza, beer and the opportunity to begin interpreting data from tens of thousands of mPower users. That event reflects another trend in crowdsourcing — the rapid spread of biomedical hackathons. These are designed to bring experts from different disciplines face to face. The Hacking Medicine initiative at the Massachusetts Institute of Technology (MIT) in Cambridge, for instance, has so far hosted almost 50 such events, teaming up engineers and data scientists with clinicians in 1- or 2-day events that are meant to quickly and iteratively work towards initial solutions to a host of health-care problems.

Among early results is an infant-resuscitation device for use in developing countries. The Ugandan paediatrician who first presented the problem has now taken the device into clinical trials in his country. The MIT initiative has helped to spark similar gatherings in places such as India and Uganda, led by the Consortium for Affordable Medical Technologies at

Massachusetts General Hospital in Boston.

Bringing researchers with varied expertise and skills together in one physical location can accelerate research. The BioMed X Innovation Center in Heidelberg, Germany, has gone further with what co-director, and biologist, Christian Tidona describes as an “outcubator”. Researchers compete not to come up with the best solution, but for the chance to try.

BioMed X begins by posting a very specific problem from one of its sponsors online. This could be exploring a new drug target or an area of treatment new to the sponsor. These requests typically get 400–600 responses from around the world. BioMed X picks 15 of the most promising concepts submitted and brings their creators to Heidelberg, where they form teams for an intense 5-day competition. The winning group then tackles the problem in two- to four-year fellowships in Heidelberg.

One of the first teams to go through the four-year exercise — made up of researchers from Germany, Slovenia and Egypt — created bioinformatics tools for designing highly selective inhibitors of kinases, proteins that play a part in many diseases. The sponsor, Merck, bought the intellectual-property rights and then licensed them back to the team, which formed a start-up company to develop the technology.

### RULES FOR THE FIGHT

The benefits for research are clear, but what is it that drives participation in crowdsourced competitions? When a challenge is centred in a researcher’s field, typically the greatest incentives to participate are the chance to publish a paper in a top journal and to network with peers, organizers say.

But often the entrants are not the usual suspects. “They’re also gadgeteers, basement inventors and weekend engineers,” says Bingham. “It’s not a bunch of French-literature majors that are solving our chemistry problems, but it might be physicists or intellectual-property attorneys or biologists.” Even in competitions with cash prizes, “at the end of the day, cash is often a scorecard, not a paycheck”, he says. Challenges would be “a silly way to make money”, says Goldbloom. The main draw for participants is what originally led him to found Kaggle — the desire for “access to interesting data sets and interesting problems”.

For medical firms, the challenges often provide a relatively quick and inexpensive way to solve tricky problems, Bingham says. At the same time, he points out, “in order to bring a product to market, they usually have to solve a thousand problems of equal complexity”. For all concerned, “the wisdom of crowds works beautifully in a great percentage of the cases”, says Stolovitzky. “We’re seeing a lot more buy-in for these challenges. If you can multiply the number of people, you can accelerate the research.” ■

**Eric Bender** is a science writer based in Newton, Massachusetts.





## COMPOUND SCREENING

# Fresh hunting ground

*In the search for novel therapies, drug developers have begun crowdsourcing molecules.*

BY ANNABEL MCGILVRAY

Across the world, more than 15,000 chemical compounds are concocted every day, most of them in university laboratories. Until recently, the majority ended up in deep freeze — or on the rubbish heap.

“When I was a PhD student I made 300 molecules,” says Matt Cooper, a chemist at the University of Queensland in Brisbane, Australia. “Three years’ work, completely new molecules, all weird, all handcrafted. I remember having each of them labelled and bottled, but basically they went in the bin.”

More than 20 years later, Cooper is director of one of the world’s largest molecular-compound-screening programmes: the Community for Open Antimicrobial Drug Discovery (CO-ADD). Frustrated by the waste of novel molecules, any one of which might hold the key to life-saving treatment

for antibiotic-resistant bacterial infection, malaria or tuberculosis, Cooper wants to put these under-appreciated creations to use.

CO-ADD is collating and screening vast quantities of small molecules in an effort to help spur the development of powerful new antibiotics. Funded by the University of Queensland and Britain’s Wellcome Trust, the programme has already received 40,000 compounds for screening since its launch in early 2015. Of these, around 26,000 have undergone high-throughput screening, an automated procedure that can test the biological activity of thousands of compounds a day and an established part of the early drug-discovery process.

The compounds are tested for antimicrobial activity against the most dangerous hospital-acquired, antibiotic-resistant infections *Escherichia coli*, *Klebsiella pneumoniae*, *Acinetobacter baumannii*, *Pseudomonas aeruginosa* and

methicillin-resistant *Staphylococcus aureus* (MRSA), as well as the common causes of fungal infections *Cryptococcus neoformans* and *Candida albicans*. The programme is working with 88 groups from 26 countries and has a further 300,000 compounds on the way, including those from what will be the first nationwide antimicrobial screening project to be coordinated by France’s National Centre for Scientific Research.

It is an initiative born of necessity at a time when profit-driven antibiotic research is floundering, says Cooper. “We think one of the reasons the industry is not productive any more is because we have been looking in the wrong chemical space.” He thinks that new drugs, including antibiotics, are likely to come from outside the conventional chemical structures held in the vast collections of pharmaceutical companies.

To see whether the molecules created in academic settings are more fruitful than the



## SEARCH HISTORY

Compound-screening programmes are relatively recent creations, and cover a range of therapeutic areas. They also vary in their processes, goals and legal requirements. The data made available to *Nature* suggest that, the more open the terms of the agreement, the more willing are external researchers to take part.

Initiative	Launch	Compounds screened	Assay focus	Intellectual property	Data
Community for Open Antimicrobial Drug Discovery (CO-ADD)	2015	Around 26,000; 88 organizations submitting	Antimicrobial	Remains with submitter	Data posted online by CO-ADD after 18 months
Eli Lilly: Open Innovation Drug Discovery (OIDD)	2009	70,000; 345 organizations submitting	Endocrine, cardiovascular, neurological, oncological and tubercular	Remains with submitter	Data are shared with submitter immediately
AstraZeneca: New Molecule Profiling	2014	Undisclosed	Cardiometabolic, oncological, neurological and respiratory	Remains with submitter	Data are shared with submitter after a year
Leo Pharma: Open Innovation	2015	150; 15 organizations submitting	Dermatological	Remains with submitter	Data are shared with submitter
Merck: Open Compound Sourcing	2011	Undisclosed; 30 organizations submitting	Oncological, immuno-oncological and immunological	Transferred to Merck when the compound is accepted for screening	Data become the property of Merck and are not automatically shared

commercial compound libraries, Cooper and his colleagues began to screen compounds put forward by an academic group for efficacy against the selected bacteria and fungal pathogens. Compared with the biological activity of commercial libraries, the academics' compounds had a 20–30 times higher hit rate. "Academics make compounds for a whole range of reasons," says Cooper. "They are more eclectic and more diverse."

CO-ADD is now leading the way in tapping this diverse pool of potential drug leads by making it as easy as possible for chemists to share their creations and uncover potentially valuable medical uses. "We make no claims on intellectual property, so all the results and all of the intellectual property rests with the provider of the compounds. There are no strings attached at all," says Cooper. "We have become the first group to take open access to the *n*th degree and really reach out to people all around the world — crowdsourcing our molecules."

In an era of low productivity in early drug discovery, medicinal chemists in the corporate realm have reached a similar conclusion to Cooper — the lack of chemical diversity is affecting early drug development. In late 2015, pharmaceutical giants AstraZeneca and Sanofi took the unusual step of exchanging 210,000 proprietary compounds to boost the variety of compounds in their collections. And, like CO-ADD, some big pharma are inviting external researchers to put forward their compounds to determine whether they may have some medical application. In a world in which knowledge is fiercely protected, the new approach is prompting ongoing adjustments by everyone involved. Academic chemists are becoming more open about their research, and companies are now more willing to share intellectual property to encourage researchers to participate. This open-innovation approach is young, and drug development is notoriously slow — a decade at best in most cases — so it may be years before a blockbuster drug results from this process. But the platforms are

showing promise, which is sparking excitement among the enthusiastic pharmaceutical executives who are driving them forward.

## CHEMICAL DIVERSITY

Screening external compounds is a ready-made source of innovation for big pharma and requires little investment, given that the companies already have the sophisticated screening technology in place. "It gives us access to compounds that are really different to the types of compounds that we have in our collection," says Garry Pairaudeau, head of external sciences at AstraZeneca in Cambridge, UK. "A lot of the academic chemistry that is going on is quite unusual and quite specialized and would be difficult for us to reproduce in our lab."

Merck KGaA, based in Darmstadt, Germany, Leo Pharma and Eli Lilly are also hosting compound-screening platforms that are designed to engage external researchers (see 'Search history'). The corporate platforms ask academic scientists or those associated with biotechnology start-ups to first submit their compounds for computer analysis. The structure of the compound is not shared with the drug giants at this stage, ensuring that the original researchers' intellectual property is protected. This analysis is carried out to determine whether the molecule meets the company's threshold for further examination. Compounds can be rejected because they are too reactive or unstable, or too similar to previously screened compounds. In the programmes run by AstraZeneca and Leo, this initial analysis is performed by a third party as an added precaution to protect confidentiality.

If compounds meet the various thresholds, the external researchers then provide a sample (typically 1–5 milligrams) for further investigation. It is at this point that the molecular structure is shared, and agreements that outline the obligations and expectations of the various parties, as well as details of where the intellectual property rights lie, are signed. With the exception of Merck, the intellectual property remains with the external researcher.

The results of the screening are then either shared with or transferred to the external researcher — again, except for the Merck programme, which provides no biological information unless an agreement to collaborate further is signed.

Each company then has a different procedure if it decides that it wants to develop the compound. Eli Lilly asks only for the right of first approach if a compound shows promise. But the company has no legal recourse if a researcher takes a molecule elsewhere following the screening. "We hope that when somebody sends a compound to us, that we have the first right of approach," says Marta Piñero-Núñez, head of Lilly's Open Innovation Drug Discovery (OIDD) platform, based in Indianapolis, Indiana. "We hope that by working with them and engaging in the research with them that we become a preferred partner." It's a similar story for Leo and AstraZeneca.

Merck's approach is more proprietorial. The standard agreement, signed when a compound is accepted for screening as part of its Open Compound Sourcing initiative, sets out that the arrangement must remain confidential and gives the company an exclusive three-year licence to develop the compound. Merck pays €200 (US\$228) for the privilege, with a promise to pay a further €5,000 if the company files a patent application that includes the compound within 5 years.

In all cases, the agreements at this point can become complex and, unlike the simpler no-intellectual-property approach of CO-ADD, require significant legal insight. "These things can take a long time," says Niclas Nilsson, head of open innovation at Leo Pharma in Ballerup, Denmark, "which only further highlights the need for speedier and less-difficult processes to explore new collaborations".

These legal hurdles may be one reason why the response to some of the corporate screening platforms has been relatively weak compared with the CO-ADD programme. Merck has signed only around 30 contracts to screen small groups of compounds since

the initiative's 2011 launch. In the seven years that the less prescriptive Lilly OIDD platform has been operating, the company has screened 70,000 molecules selected from 400,000 submissions, and at the start of 2015 had 345 groups participating. The screenings have been done at a consistent rate of about 500 compounds per month, but according to Piñero-Núñez, the company has the capacity for thousands more. Meanwhile, Leo has screened just 150 compounds submitted by 15 partners in the first 12 months of its Open Innovation programme — a long way from the roughly 26,000 that have been assayed by Cooper and his CO-ADD colleagues during that time. Cooper attributes the variation to concern about the motivation behind the programmes. "People are probably more likely to engage with a not-for-profit," he says. "They feel more comfortable that we're doing it for the right reasons."

Aware of the challenge of fostering participation, Nilsson set out to devise a screening programme that would be attractive to both small industry and academia. Nilsson says, "Our approach was based on interviews with biotech and university researchers to determine how low the bar would have to be set in order to persuade them to take part and share their creations". He then negotiated with the company's legal and business development arms to determine how barriers to submission could be reduced.

Researcher feedback drove Nilsson's team to further adapt the conditions of the programme. Initially, they arranged to share the data with participants, giving both parties rights of access and ownership. But it quickly became clear that this would be difficult for small biotechs to agree to — if they don't own the data, they can't use them as leverage when negotiating with others. "So now we transfer the ownership of the data back to the external partner. It means that they can increase the value of their assets by using our services and then they can walk away if they want to. That is, of course, our risk," says Nilsson.

Although the numbers are small, Nilsson says that the fledgling programme has already generated positive responses and the company is in the process of negotiating with one contributor on a future collaboration in relation to a treatment for the skin condition psoriasis. "They submitted compounds that are targeting new proteins that we didn't even know existed. So, not only are they providing new chemistry, but they also open up a new science that we didn't know could develop," Nilsson says. "That is exactly what we're looking for with the open innovation — something new that we couldn't predict ourselves."

## PROMISING PROGRESS

It is too early to see the fruits of this open-innovation approach in the marketplace. But there are signs that this increased engagement



Ana Cristina Parra Rivera, a student of Doug Frantz at the University of Texas, works remotely for Eli Lilly.

with researchers and their molecules is having an effect. Doug Frantz, a chemist at the University of Texas at San Antonio, has been contributing to the Eli Lilly compound-screening programme for five years. His laboratory has submitted around 500 compounds and recently became the first OIDD contributor in the programme's history to reach an agreement to progress a molecule to trials in animals. In the context of drug development, this is more than one-third of the way towards

**"A lot of the academic chemistry is quite unusual and quite specialized."**

clinical use. The progress of the potential chronic-pain drug (an antagonist for a form of sodium channel called NaV 1.7) is a milestone for him as well as for Lilly. "It's just a fantastic opportunity," Frantz says, "to be able to explore the biological activity of my molecules, rather than just making them and sticking them in the freezer and publishing a paper. I can take that data and I can take the compounds and publish or put it into a grant application if I want."

Frantz and his students can keep tabs on the assays that their compounds are undergoing and any results through a secure online account. Many of their compounds have shown activity in these assays, and there are a few particularly promising compounds that they are considering adapting to meet Lilly's threshold for further collaboration. As well as the NaV 1.7 antagonist, they have a second contract with Lilly that relates to a possible treatment for schizophrenia (an antagonist of the nervous-system receptor mGluR2). This has involved one of Frantz's students working over the Internet to create compound analogues in Lilly's automated-synthesis lab.

The University of Texas at San Antonio is one of 255 universities that are contributing compounds for screening with Lilly. Piñero-Núñez says that, initially, "people would get up on their seats and say that we were

crazy", but now many in the academic community as well as industry are embracing external compound screening. The results of the screening of compounds produced by external researchers by Lilly have been comparable with, and sometimes better than, the 1–5% hit rate of internally produced molecules. "Even if the molecules don't advance, they introduce starting points and they contribute to each project," Piñero-Núñez says. The screening platforms are also a way to save money — AstraZeneca has put the cost of the internal production of the 210,000 compounds that it received from Sanofi last year at around \$25 million.

After about a year of screening, the CO-ADD project has uncovered 128 hitherto unknown potential antibiotics that have passed tests for cell toxicity. On the basis of these results, the group is already talking to health-care-research funders, including the US-based Bill & Melinda Gates Foundation and the Wellcome Trust, about extending the programme. By the end of 2020, Cooper hopes to have screened 1 million compounds. "When we do these things collectively and collegially, they're of incredible value to society," he says. "This is like a molecule bank — we want to use this to screen for new compounds against malaria, against tuberculosis, against other pathogens and threats to world health."

CO-ADD seems to be well on its way to sourcing those 1 million molecules, but it remains very early days for this kind of open-innovation approach. Already, there is new engagement between big pharma and academia, and more diversity in compound collections and novel research leads. But only time will tell if drug developers can capitalize on this early promise. With only one compound thought to have made it to animal trials, these programmes still have a long way to go if society is to feel their true effect. ■

**Annabel McGilvray** is a freelance writer based in Sydney, Australia.





A child in Tanzania is treated for malaria — a disease that could benefit from an open approach to medical research.

## TROPICAL DISEASE

# A neglected cause

*A more open approach to combating tropical diseases may help to overcome a pharmaceutical market failure.*

BY LUCAS LAURSEN

He didn't know it at the time, but when chemist Matthew Todd posted a request for help on The Synaptic Leap, a website devoted to open-source biomedical research, he was sowing the seeds for a rivalry between an open initiative and a contract-research organization hired by the World Health Organization to reach the same goal.

The aim of both projects, run in 2010, was to produce a safer, low-cost version of praziquantel, a treatment for the tropical parasitic infection schistosomiasis. Up until that point, the treatment contained two enantiomers (mirror-image versions of the molecule that have slightly different properties) of praziquantel. One enantiomer has no effect on the parasite, but gives the drug a bitter taste. Eliminating this undesirable form could reduce side effects and help more patients to complete their treatment. The pure drug needed to be affordable. Todd, who is at

the University of Sydney in Australia, thought that an open project was the best way to achieve this. "Open is very well-suited for neglected diseases," he says. "The pay-off of secrecy is not very large."

This is because the potential buyers of drugs targeted at rare and neglected diseases, such as schistosomiasis and dengue fever, are often unable to pay the prices that large pharmaceutical companies must charge if their investments are to pay off. Occasionally, these companies have made investments that have no hope of direct commercial return, such as the decision in 1987 by Merck and Co., based in Kenilworth, New Jersey, to offer Mectizan (ivermectin), a drug for onchocerciasis or river blindness, free of charge. But too many of these decisions would risk a company's survival. Although governments have often intervened to promote and fund research in areas with the highest financial risk, diseases that affect the poorest countries or smallest subpopulations

remain undertreated and under-researched.

The race to produce a single-enantiomer version of praziquantel finished in late 2010 in an effective tie when both Todd's open group and the contract-research organization posted their different but comparable solutions online. Half a decade later, Todd is aware that his method is still being used to produce the active enantiomer of praziquantel in various labs around the world. Todd points to the work as one of the success stories of open-source medical research. The project matched the results of the conventional approach, but, by collaborating with researchers previously unknown to Todd's team, he says that the project did it faster than it ever could have done in secret.

The research community is embracing the open approach to find treatments for diseases that have little potential to make a profit. By collaborating, drug companies, non-governmental organizations and governments can share the load. Partnerships to handle development and

STEPHANIE AGLIETTI/AFP/GETTY



distribution of treatments for neglected and tropical diseases have already begun, as have efforts to find ways to share data on early-stage research to avoid duplication. And although open-source projects, such as Todd's, remain the exception rather than the rule, they, too, are gaining popularity.

### FASTER SCIENCE

Around the turn of the millennium, a number of organizations were set up to connect industry, government and non-governmental organizations, and to promote the sharing of knowledge about tropical and neglected diseases — Medicines for Malaria Venture (MMV), the TB Alliance and the GAVI Alliance. Pharmaceutical companies, they argued, had compound libraries and dormant intellectual property that, for areas in which they could not expect to reap significant profits, could be shared without threatening their financial prospects.

In 2010, pharmaceutical companies GlaxoSmithKline and Novartis, and St. Jude Children's Hospital in Memphis, Tennessee, publicly released a database, through MMV that contained 20,000 compounds that had shown some antimalarial activity. These 'hits' are an early step on the long path to designing new drugs (see page S65) and are usually proprietary. Researchers began to also ask for samples that they could work with, so, the following year, MMV began offering a selection of the compounds by post. This 'malaria box' offered researchers a common starting point and the organizers hoped that it would accelerate research.

The box meant that researchers did not waste valuable time creating duplicates of compounds already produced by someone else. "People could really focus on understanding novel chemical series," says Jeremy Burrows, head of drug development at MMV based in Geneva, Switzerland. And thanks to MMV's coordination, researchers can divide up the analysis to understand the existing pool of compounds, rather than work in secret and risk overlapping efforts.

Companies have also begun to share their intellectual property through large international systems, such as the Re:Search project, established in 2011 by the World Intellectual Property Organization (WIPO) to drive innovation around neglected diseases, malaria and tuberculosis. Data and compounds, for example, are made available through research agreements that specify certain requirements, such as assurance that access to any end product is affordable in the 49 least-developed countries. As of March 2016, the programme had 103 members operating under 100 research agreements. Five active antimalarial projects involve compounds

in pre-clinical development — one step before human clinical trials — and a WIPO-initiated dengue-fever project is already putting a repurposed compound through clinical trials.

### OPEN REVOLUTION

When business specialist Henry Chesbrough of the University of California, Berkeley, popularized the term open innovation in 2003, his definition was a broad one. It included industrial adoption of academic practices, such as publishing results in journals, and collaborations, within which a select few partners would share information with each other, but maintain secrecy with the wider world. Todd's approach with open-source science was more radical.



The female (green) and male parasites that cause the tropical disease schistosomiasis.

"In drug discovery and development, you have a spectrum," says Burrows. At one end are pharmaceutical firms going about their work in secret, he says. At the other is open source. This model, whereby information is made freely available to a wider community and not just the people directly involved in the collaboration, originated in software development. Central to Todd's project was the decision to make the team's laboratory notes available to the world while they worked. Although this meant that any other teams working towards the same objective were able to see exactly what they were doing, it also opened them up to external help. "The people contributing really knew what they were doing," says Todd. Not only does he say that their input accelerated the research, but also that timely suggestions to change direction kept them away from expensive blind alleys. "We could have wasted the whole grant on something unproductive," he says.

Sharing lab notes can also prevent duplication of effort, says Katy Graef, associate director of BIO Ventures for Global Health in Seattle, Washington. In one case, Graef says, the non-profit organization learned that researchers at Saint Louis University in Missouri were planning to screen inhibitors of the enzyme METAP-1 for antitubercular activity, a test that GlaxoSmithKline had already performed. BIO Ventures put

the two parties in contact, and the Saint Louis researchers were able to redirect their research, instead of wasting an estimated three months.

Since his first foray into open-source research, Todd has developed a bigger project: Open Source Malaria. The aim of the project is to find new medicines for malaria. So far, the team has explored several groups of promising compounds, including those released through MMV, and it is currently exploring the potential of a set of molecules that originated with the drug giant Pfizer.

Anyone can contribute to the project, and with all data and ideas posted publicly, this is perhaps the most open approach yet to tackling a tropical disease. The project is part of an emerging trend in drug research. The Open Source Pharma conference, sponsored by the Open Society Foundations and the Rockefeller Foundation, was held in 2014 and 2015. And since 2008, the Indian government-funded Open Source Drug Discovery project has been conducting crowdsourced tuberculosis research with the aim of expanding to other neglected diseases. The project has identified potential molecules for fighting tuberculosis and built infrastructure for coordinating remote efforts on a tight budget.

The global burden of neglected tropical diseases, as defined by the World Health Organization, as well as malaria and tuberculosis is around 5.5% of the total number of healthy years lost to disability, poor health or early death. But research and development spending on neglected diseases in 2010 was only around 1% (about US\$2.4 billion) of global-health research spending. And only 4% of therapies registered between 2000 and 2011 were indicated for neglected diseases. The open research initiated in the past decade could begin to address these imbalances.

Not all initiatives to tackle neglected diseases need to be run as publicly as Open Source Malaria to make an impact. The Wellcome Trust, for instance, began offering grants for multi-centre, transnational collaborations in 2015, inspired by the ad hoc partnerships that sprang up between labs during the Ebola crisis. "There are a lot of different models out there," says Graef, "and it's good that we have these different approaches to keep things complementary."

Burrows shares her optimism, and points to four promising compounds that emerged from MMV's open work and are included in candidate drugs now in clinical testing. It will be a few more years before any open-origin treatment for neglected diseases makes it from the sharing stage all the way through to commercialization, but that is where everyone, from non-governmental organizations to the largest pharma firms, is placing their bets. "Everyone is aware that opening things up makes them more efficient," Todd says. ■

Lucas Laursen is a freelance journalist based in Madrid.

## PERSPECTIVE



# Science is still too closed

Open initiatives are promising, but we have much further to go if research data are to be as publicly accessible as they should be, says **Aled Edwards**.

To paraphrase Joy's Law, no matter where you work, most of the smartest people are somewhere else. This principle, coined by co-founder of Sun Microsystems Bill Joy, could also apply to the best data or the most cutting-edge technologies. Providing open access to these distributed assets would accelerate science and innovation. But, despite the promise that open access holds, it has so far proved difficult to implement.

Few would dispute that sharing science accelerates the rate of discovery. The fact that open science also boosts innovation in industry is less well appreciated. Economist Heidi Williams at the Massachusetts Institute of Technology in Cambridge retrospectively analysed the commercial activity that flowed from two large sets of sequenced human genes (H. L. Williams *J. Polit. Econ.* **121**, 1–27; 2010). One group was made up of sequenced genes that for a period of time had been available only under commercially restrictive terms. The genes in the other group had always been in the public domain. Over a 10-year period, Williams found, there was around 30% more commercialization activity from the open set of genes.

A look at the biopharmaceutical sector reveals a similar story. Both monoclonal antibody and phage-display technologies are used to identify precursors to antibody drugs. Both technologies were invented more than 25 years ago, and both have been used to discover successful medicines. But whereas the phage-display technology has been fiercely protected, monoclonal antibody technology was placed in the public domain. As of 2014, there were 47 approved monoclonal antibody drugs and only 7 derived from phage-display technology.

Although it may seem counter-intuitive, openness is good for innovation. No wonder, then, that the idea is gaining traction in biomedical research circles. 'Open' has become one of the hottest topics in boardrooms, funding agencies and the media, and dozens of initiatives have been launched under the open brand.

Open-access initiatives to make the scientific record more widely accessible are being championed by charitable and governmental organizations. Most of the largest funders now require that articles that are derived from research that they support are made freely available on the Internet after a period of time, often a year or less. These organizations are also working to ensure the publication of large data sets through initiatives that are modelled on the Human Genome Project (HGP), which released data daily and without restriction on their use.

These are positive steps, but there remains much room for improvement. Open-access publishing, for instance, is often possible only when publishers charge hefty publication fees (which are paid for with public funds) and some of the highest-impact journals continue to resist open-access policies. Few of today's open-data initiatives meet the 20-year-old standards set by the HGP — most allow data release to be delayed and let primary investigators control the data, and many place various restrictions on data use.

Indeed, few initiatives are truly open. The term open innovation, as defined in the management literature in 2003 (H.W. Chesbrough *Open Innovation: The New Imperative for Creating and Profiting from Technology*; Harvard Business Press, 2003), refers to a collaboration in which two or more companies share or license proprietary information between themselves — a far cry from true openness. More recently, companies have begun to embrace less restrictive collaborative approaches to access more ideas and technologies, including crowdsourcing projects (see page S62) and pre-competitive consortia (see page S56). However, industry-led initiatives that yield publicly accessible research are still rare.

The open-access movement has gained so much momentum that it can be tempting to believe that everything is awesome. The reality is more nuanced. Although the progress towards open access is encouraging, there is a long way to go before all scientific results are communicated in real time, at no cost and without restriction on use as a matter of course.

Considerable change is needed. Research produced in universities should be available to all, but it is not. Universities often limit access to their research output because they continue to adhere to the ideology that secrecy and patents are obligatory foundations for commercialization and innovation. The imbroglio over who owns the rights to the CRISPR–Cas9 gene-editing technology will probably emerge as another case study for how the financial interests of institutions can inhibit innovation by limiting, rather than promoting, the uptake and application of foundational technologies.

Experimental reagents and protocols should be freely available to allow researchers to reproduce experiments; this is not always the case, and even when it is, most are encumbered by legal agreements that restrict their use.

Data from clinical and genetic studies should be made available to the study participants, but they are not. In most such studies, the data are considered to be proprietary, and there is no obligation to release them to the participants of the study, much less the public.

If this is to change, I propose that society first agree on a simple, guiding principle: all scientific discoveries first constitute a public good and only second are the property of individual scientists, institutions or countries. Agree on this, and it follows that anything that impedes the sharing of discoveries — either by prolonging the time or complicating the process of disseminating scientific outputs — should be eliminated entirely. We should not be satisfied with anything less. ■

**Aled Edwards** is chief executive of the Structural Genomics Consortium at the University of Toronto in Canada.  
e-mail: [aled.edwards@utoronto.ca](mailto:aled.edwards@utoronto.ca)

DESPITE THE  
PROMISE THAT  
**OPEN ACCESS**  
HOLDS, IT HAS  
SO FAR PROVED  
DIFFICULT TO  
**IMPLEMENT.**





A researcher at the Montreal Neurological Institute and Hospital, which is making all its data public.

#### DATA SHARING

# Access all areas

*Advocates say that open science will be good for innovation. One neuroscience institute plans to put that to the test.*

BY BRIAN OWENS

In the cut-throat world of early-stage clinical development, where aggressive defence of data and intellectual property is thought to be key to amassing profits, one academic institute is opting out.

Over the next five years, McGill University's Montreal Neurological Institute and Hospital (the Neuro) in Canada will conduct a radical experiment in open science. It will make all results, data and publications from its research free to access, will require collaborators to do the same, and, perhaps most surprisingly, will not pursue patents on any of its discoveries.

The primary motivation for the move is to increase the pace of discovery in neuroscience — a field in which clinical progress has so

far been slow. “We think that by sharing data quickly, we’ll be able to accelerate the discovery of mechanisms and eventually new medicines,” says Neuro director Guy Rouleau.

But Rouleau acknowledges that there is also a moral argument for opening up scientific data. “We’re funded mostly by public money, so it makes sense that it be freely available.”

The Neuro’s open policy, expected to come into effect this summer, is based on five principles developed through a series of consultations with the institute’s faculty and staff. The first is that Neuro researchers will make all information about a study publicly available by the time the research is published. This requirement

will apply to all of the results — positive and negative — as well as models, algorithms, reagents and software.

Second, all data and resources generated through new research partnerships — whether they be with companies, institutes or other universities — must follow the same rules. Third, the institute’s biobank, which contains tissue samples and brain-scan data, will be opened up (although the institute may charge users a small fee to cover operating costs).

The fourth principle is that the institute will not pursue any intellectual-property protections for research discoveries. And the fifth is a commitment that, although the institute will not support activities that undermine these open-science principles, it will respect its researchers’ autonomy. In practice, this means that a researcher could pursue a patent on their work, but the Neuro would not pay any of the fees or help with the paperwork.

By sharing data and results early and often, scientists should get a better idea of what is going on elsewhere in their field, and avoid exploring blind alleys that others have already rejected. This is particularly important in neuroscience, says Rouleau, where progress is slowed by both the vast complexity of the brain and the heterogeneity of neurological diseases.

Having a solid understanding of brain mechanisms is important for pharmaceutical companies working to develop new treatments, says Viviane Poupon, the institute’s director of partnerships and strategic initiatives. “Without mechanisms, they’re just fishing randomly,” she says. “We’re trying to help diminish the 95% failure rate of drug candidates targeting the central nervous system.”

How much the principles of open science really can speed up innovation is a question that the Neuro hopes to help answer. Richard Gold, a researcher at McGill who studies the use of intellectual property in the life sciences and helped the Neuro to design its open principles, will be monitoring the institute’s performance. He will then compare it with that of similar institutions that are not pursuing openness to see whether the Neuro’s decision leads to better scientific outcomes.

The specific metrics he will use are yet to be decided, but he suggests that he will be looking at how closely the institute follows its own principles, whether it inspires others to follow suit, whether people make use of its open resources, and how successful it is at attracting funding and staff.

Open-science advocates hope that greater access to information will also help to solve a larger problem facing science than the slow translation of research into products: the fact that many high-profile results have proved impossible to replicate. “We’re facing a credibility problem. Not a month goes by without some field of science being rocked by scandal,” says Björn Brembs, a neurobiologist and open-science advocate at the University of Regensburg.

➔ **NATURE.COM**

To read more about open science, visit:  
[go.nature.com/rpucxz](http://go.nature.com/rpucxz)





Guy Rouleau, director of the Neuro, favours open science.

in Germany. “Of the five or six large-scale replication studies that I know of, none of them confirm more than 50% of results.”

It's no wonder, then, that the pipeline of new treatments has slowed to a trickle. “If the pre-clinical work is not replicable, then you can't make a drug out of it,” says Brembs. Mike Ehlers, the chief scientific officer of Pfizer's Neuroscience and Pain Research Unit in Cambridge, Massachusetts, agrees. “The number of key findings that we are able to robustly reproduce is not what I would want to see.”

Christof Koch, president of the Allen Institute for Brain Science in Seattle, Washington, thinks that having all the data from every experiment freely available will go a long way towards alleviating this problem. “If science wants to overcome this crisis, this lack of reproducibility, we have to practice what we preach, and practice open science,” he says.

### GROWING FAMILY

The Neuro is joining a growing movement towards the free sharing of scientific data and results. Some large-scale projects, such as the international Human Genome Project, have freely shared all of their data, and many charitable funders, such as the Bill & Melinda Gates Foundation, require the researchers they support to make their data and published results freely available.

The Allen Institute, launched in 2003 by Microsoft co-founder Paul Allen, has followed an open-science model “from the get-go,” says Koch. The institute offers free access to its huge gene-expression maps for mice, humans and other animals, and posts data from its research online as soon as it is ready, rather than keep it hidden from researchers at other institutions until they are ready to publish a complete paper. “All our biggest papers have been published two

to three years after the data was put online,” says Koch. “The idea that you need to hold on to your data until after you publish is not true.”

Drug companies are also embracing the idea that there are advantages to sharing in the early stages of research. “In the past 10 years or so we've seen greater movement towards multi-party collaborations,” says Gold. “Generally in areas of science that are expensive, and considered pre-competitive and high risk.”

The big pharmaceutical companies can see the value of working more openly with academic neuroscientists and with each other to mitigate those expenses and risks. “It's a very complex field,” says Hans Lindner, head of global external innovation and alliances at the pharmaceutical company Bayer in Berlin. “Combining efforts is essential to deal with complex matters.”

Much of the work is done in large groups involving multiple companies, universities and public agencies, all combining their cash, expertise and equipment, and sharing the results. “They realized that they're all spending money doing the same thing, and they could leverage that money through partnerships to do more risky work,” says Gold. This provides universities with a new source of funding and allows government agencies to steer work towards their priorities. It also benefits the companies, which see more efficient discovery of potential drug targets.

Rouleau says that his inspiration for the Neuro's open-science project came from one such large-scale collaboration, the Structural Genomics Consortium (SGC). The SGC is a collaboration between six universities and eight pharmaceutical companies from around the world. It receives funding from government agencies in Brazil and Canada, and the UK Wellcome Trust, and generates protein structures, chemical probes and antibodies to speed up the development of new drugs. “When you can get academics to share, and get industry to share, you get a hell of a lot more and better data out of that collaboration,” says Aled Edwards, the SGC's director. “It's a fantastic way to do science.”

Lindner says that open innovation is “an essential part of our R&D strategy”. In addition to taking part in the SGC and the European Union's Innovative Medicines Initiative, a public-private project to boost pharmaceutical innovation, Bayer has several crowdsourcing initiatives that provide access to Bayer compounds, and an ‘incubator’ laboratory that provides start-ups with lab space and access to its facilities. “The benefit of the open source is it may ease up the early testing of concepts, and increase the chance of an interesting concept being further developed,” Lindner says. “Ultimately, we have more shots on goal.”

### NO PATENTS PLEASE

Although open science is a growing trend, at least one aspect of the Neuro's plan seems to be unique — the vow to eschew all

intellectual-property protections on its work. It's a move that is proving popular in the open-science community, where most feel that the drive to protect every discovery has gone too far and is stifling progress. “Patents don't help drive innovation,” says Edwards. “They just get in the way most of the time.”

Ehlers is even more blunt: “Universities tend to slather IP on every finding, regardless of its potential value.”

Lindner, meanwhile, is intrigued by the ‘no patents’ approach, and sees some potential benefits. “It may ease up the initial interaction between the institute and other parties,” he says, when legal niceties such as intellectual property rights and licensing fees are usually decided.

**“Institutions waste a lot of time patenting, and most patents don't generate any money.”**

“Otherwise you have to negotiate this at the beginning, which can be lengthy and frustrating.”

However, Ehlers does foresee complications for companies that want to develop products that are based on the Neuro's work. Companies often prefer to work with protected ideas, because it gives them a way to recoup their investment. “There has to be a well-calibrated use of patents,” Ehlers says. “If there's too little protection, there's no way to capture the value.”

Rouleau says that the institute is not completely opposed to the idea of patents. If a Neuro discovery shows commercial promise, a pharmaceutical company will be welcome to take it in house and use it to develop a patentable medicine. Rouleau acknowledges that this means that the institute and McGill would then lose out if one of the Neuro's ideas became a blockbuster drug, but he thinks that risk is low. “We're working at such early stages, anything we discover will need to be taken and worked on for years — our share of any profits would likely be small,” he says.

And any loss should be more than made up for by new investments from philanthropic organizations and companies, from whom the Neuro's open concept is already attracting serious interest.

Gold agrees that passing up patents will be no great loss for the Neuro. “Institutions waste a lot of time patenting, and most patents don't generate any money,” he says. “This gets the university out of the business of business, and back into knowledge generation.”

Besides, the institute's top priority should be to do its best for the people relying on it to help cure their illnesses, not to make money, says Rouleau. “Blocking other people from working with our findings is not in the best interest of patients,” he says. ■

**Brian Owens** is a freelance science writer based in St. Stephen, New Brunswick.

# Industry–academia collaborations for biomarkers

*Khusru Asadullah<sup>1,2</sup>, Andreas Busch<sup>1</sup>, Matthias Gottwald<sup>1</sup>, Petra Reinke<sup>2</sup> and Lilla Landeck<sup>2,3</sup>*

Several types of collaboration are being pursued to identify, validate and apply new biomarkers. Here, we highlight examples of such initiatives and discuss the challenges, approaches to address these challenges and key factors for success.

Biomarkers are becoming increasingly important for predicting disease prognosis, enabling personalized therapy (also known as precision medicine) and detecting early therapeutic and adverse responses to drugs. However, the identification, validation and application of biomarkers is challenging, with several aspects, including understanding of the biology of the biomarker and its relevance to disease, the technological characteristics of the assay used for biomarker measurement and the potential regulatory requirements. The challenges are particularly demanding when co-developing a drug and a companion diagnostic based on a biomarker. In many cases, the breadth and complexity of research needed mean that collaborations are vital. This article highlights examples of such initiatives and discusses key factors for success.

## Public–private partnerships

Substantial public funding has become available for biomarker research in recent years. Public–private partnerships (PPPs) such as the European Innovative Medicines Initiative (IMI), the US [Biomarkers Consortium](#) and others (see [Supplementary information S1 \(table\)](#)) offer the opportunity to create a critical mass of partners in a specific area. Usually, several companies and academic institutions join forces to address an issue that would be hard or even impossible to tackle alone.

For example, in 2006, the US Food and Drug Administration (FDA), the US National Institutes of Health (NIH) and the Pharmaceutical Research and Manufacturers of America (PhRMA) founded the Biomarkers Consortium<sup>1</sup>. To date, the consortium has launched 18 projects in four disease areas — cancer, inflammation and immunology, metabolic disorders, and neuroscience — and nearly half of these projects have been completed. The consortium has made notable contributions to accelerate the drug development process and to enhance regulatory decision-making. For example, the establishment of a set of evidence-based

recommendations on new interim end points supported the FDA's efforts to approve new agents for community-acquired bacterial pneumonia and acute skin infections<sup>1</sup>. Another important project, the I-SPY 2 trial of potential drugs to treat breast cancer, has illustrated how innovative adaptive trial designs may substantially reduce the time and cost of getting new therapies on the market<sup>1</sup>. The [Predictive Safety Testing Consortium](#), which originated from the FDA's Critical Path Initiative and has been collaborating for several years with the Biomarkers Consortium on a kidney safety project, has qualified seven biomarkers of nephrotoxicity for use in preclinical toxicology studies and continues to work towards additional safety-biomarker qualifications<sup>2</sup>. Work by another consortium originating from the Critical Path Initiative, the [Coalition Against Major Diseases](#), has led to the qualification of low baseline hippocampal volume as assessed by magnetic resonance imaging as an enrichment biomarker for clinical trials in pre-dementia stages of Alzheimer disease<sup>2</sup>.

Biomarker-related programmes are also being pursued through the IMI. For example, the [SAFE-T](#) project has evaluated more than 150 biomarkers for drug-induced liver, kidney and cardiovascular toxicity on the basis of protocols defined upfront together with the European Medicines Agency (EMA) and the FDA. The 20 most promising biomarkers are just entering clinical validation. Another interesting example is the recently constituted IMI consortium [CANCER-ID](#), which aims to evaluate the clinical utility of different technologies for the enrichment, isolation and analysis of circulating tumour cells (CTCs), circulating tumour DNA and microRNAs. These biomarkers could support the concept of a 'liquid biopsy', which may allow for longitudinal sampling when standard biopsies are not available or pose a considerable risk to the patient. The 36 CANCER-ID partners from academia, hospitals, small-to-medium-sized enterprises, non-profit organizations, and diagnostics and pharmaceutical companies are now working on defining standards

<sup>1</sup>Bayer Global Drug Discovery, Müllerstrasse 178, 13353 Berlin, Germany

<sup>2</sup>Charité University Medicine Berlin, Charitéplatz 1, 10117 Berlin, Germany

<sup>3</sup>Potsdam General Hospital, Charlottenstrasse 72, 14467 Potsdam, Germany

Correspondence to K.A.  
e-mail: [khusru.asadullah@charite.de](mailto:khusru.asadullah@charite.de)  
doi:10.1038/nrd4727

Published online  
30 October 2015

for CTC identification and counting as well as defining the standard operating procedures for pre-analytical sample handling and analyses.

PPPs also facilitate the testing of novel approaches, such as systems medicine for the identification of biomarker patterns that could improve patient stratification. One example is the IMI project [OncoTrack](#), which is using novel computational modelling approaches to analyse deep sequencing data from >200 patients with colon cancer.

Another European consortium, known as [BIO-DrIM](#) (Biomarker-Driven IMmunosuppression), will apply recently identified biomarkers to help manage immunosuppression after solid organ transplantation. In several clinical trials that are supported by diagnostics and pharmaceutical companies, biomarkers will be used for the first time as decision criteria to guide personalized immunosuppression, which could allow the minimization of or even weaning from drugs without harming the kidney and liver allografts. The methodological implementation of these biomarkers is finalized, and up to 1,000 patients will be enrolled in trials.

Crucial factors for the success of such complex consortia include: finding a good balance between critical mass and manageable size; involvement of both diagnostics and pharmaceutical industry partners; early alignment on the goals between the public and the private partners; agreement on a suitable intellectual property framework before the project starts; and professional project management.

### Open innovation

The use of open-innovation models, also known as 'crowd sourcing', for biomarker discovery and validation is still in its infancy<sup>3</sup>. A pioneering example is provided by [InnoCentive](#), which was established in 2001 to match problems in a wide variety of fields with potential solution providers, using an online platform on which challenges are posted by organizations. For example, a challenge with a US\$1-million prize was set up in 2006 to find a biomarker for amyotrophic lateral sclerosis (ALS) by [Prize4Life](#), a non-profit organization established to accelerate the discovery of ALS therapies. This challenge attracted ~1,000 teams from ~20 countries, and the prize was ultimately awarded in 2011 for the development of electrical impedance myography to assess the progression of the disease. The results of another crowd-sourcing challenge set up by Prize4Life and the DREAM Project, which involved the use of clinical trial data from ALS patients to develop algorithms for the prediction of disease progression, have recently been published<sup>4</sup>.

Another example is Bayer's [Grants4Targets](#) (G4T) initiative<sup>5</sup>, which was established in 2009 as a mechanism for academic institutions to apply for grant support to pursue ideas on novel drug targets, and which was expanded to include biomarkers in 2011. In addition to financial support, Bayer provides specific knowledge about drug discovery and development. More than 1,000 applications have been received so far; 126 (~16%) have been for biomarkers, of which 18 have been approved for grants. These are focused on biomarker identification in

cardiological, oncological and gynaecological indications, often using novel technologies. As the G4T programme is quite young, it is too early to judge its ultimate success in the identification and validation of biomarkers. However, on the basis of our experience so far, this type of collaboration attracts academic groups and offers the chance to gain access to new partners. Key prerequisites for success are: fast and efficient processing of the requests; a low bureaucratic burden to generate and grant the proposals; and intensive and direct personal contact with the scientists in the academic institutes after grant approval.

### Concluding remarks

Collaboration between academia, the diagnostics industry and the pharmaceutical industry is particularly important for successful biomarker identification, validation and application, and requires substantial resources and complementary skills. We expect the high number of PPPs in this area will continue to increase, in part owing to the growing availability of public funding.

Different models may be best suited for different stages of the biomarker identification and validation process and for different kinds of biomarkers. For the initial exploration of a new biomarker candidate or a novel technology, an open innovation approach may be particularly suitable, as a lot of creativity, flexibility, and freedom is needed at this early stage. For the subsequent clinical validation of a biomarker to the standards required for acceptance by regulatory authorities, larger consortia may be necessary given the extent and complexity of the work involved.

Each category of collaboration has its own needs to be successful. However, there are some general factors. Organizations and people will only actively engage in collaboration when the benefit they derive is greater than the effort and time it takes to collaborate. Organizations also need to foster their competencies in relationship- and conflict-management and skills in working across the diverse cultures in academia and industry. Finally, real internal commitment from the organizations involved is essential for any collaborative effort to be successful.

1. Wholley, D. The Biomarkers Consortium. *Nat. Rev. Drug. Discov.* **13**, 791–792 (2014).
2. Stephenson, D. & Sauer, J. M. The Predictive Safety Testing Consortium and the Coalition Against Major Diseases. *Nat. Rev. Drug. Discov.* **13**, 793–794 (2014).
3. Lessl, M. *et al.* Crowd sourcing in drug discovery. *Nat. Rev. Drug. Discov.* **10**, 241–242 (2011).
4. Küffner, R. *et al.* Crowdsourced analysis of clinical trial data to predict amyotrophic lateral sclerosis progression. *Nat. Biotech.* **33**, 51–57 (2015).
5. Dorsch, H. *et al.* Grants4Targets — an open innovation initiative to foster drug discovery collaborations between academia and the pharmaceutical industry. *Nat. Rev. Drug. Discov.* **14**, 74–76 (2015).

### Acknowledgements

The authors would like to thank T. Schlange, M. Lessl, H. Dorsch (all at Bayer, Germany), and H.-D. Volk (at the Berlin-Brandenburg Center for Regenerative Therapies, Charité, Berlin, Germany) for helpful discussions, and P. Carrigan for critical reading of the manuscript.

### Competing interests statement

The authors declare [competing interests](#): see Web version for details.

### SUPPLEMENTARY INFORMATION

See online article: [S1](#) (table)

ALL LINKS ARE ACTIVE IN THE ONLINE PDF



## OPINION

# Hit and lead criteria in drug discovery for infectious diseases of the developing world

Kei Katsuno, Jeremy N. Burrows, Ken Duncan, Rob Hooft van Huijsduijnen, Takushi Kaneko, Kiyoshi Kita, Charles E. Mowbray, Dennis Schmatz, Peter Warner and B. T. Slingsby

**Abstract** | Reducing the burden of infectious diseases that affect people in the developing world requires sustained collaborative drug discovery efforts. The quality of the chemical starting points for such projects is a key factor in improving the likelihood of clinical success, and so it is important to set clear go/no-go criteria for the progression of hit and lead compounds. With this in mind, the Japanese Global Health Innovative Technology (GHIT) Fund convened with experts from the Medicines for Malaria Venture, the Drugs for Neglected Diseases *initiative* and the TB Alliance, together with representatives from the Bill & Melinda Gates Foundation, to set disease-specific criteria for hits and leads for malaria, tuberculosis, visceral leishmaniasis and Chagas disease. Here, we present the agreed criteria and discuss the underlying rationale.

There is an urgent need for new and more-effective drugs to treat the various diseases that take the heaviest toll on the developing world<sup>1</sup>. This can only be achieved in a cost-effective manner by implementing robust and efficient processes to develop and deliver drugs that are safe, effective, affordable and available to those who need them most.

The quality of chemical starting points (known as 'hits') for drug discovery projects is a key factor for improving the likelihood of success of clinical candidates; starting a discovery project with poor-quality hits ultimately results in increased attrition of these compounds<sup>2,3</sup>. The decisions to progress a hit into 'lead' identification (the HTL phase) and then on into lead optimization are crucial, as the downstream optimization phase may take years and require considerable financial investment. Setting the bar high by applying comprehensive, well-considered criteria for entry into lead optimization will improve overall success rates and focus resources on chemical series that stand a reasonable

chance of delivering a quality preclinical candidate<sup>4</sup>. The need to focus resources effectively is particularly important in drug research and development (R&D) for infectious diseases that affect people in the developing world, given the limited market incentives for R&D investment and the key role that philanthropic funding and public-private partnerships have in this field.

The Japanese Global Health Innovative Technology (GHIT) Fund (BOX 1) is a pioneering public-private partnership. It provides resources for research organizations in Japan to partner with international research groups and product development partnerships (PDPs) to screen and identify new hits and develop them into novel lead series for infectious diseases that affect people in the developing world. The initial areas of focus for the GHIT Fund are malaria (in partnership with the Medicines for Malaria Venture (MMV)), tuberculosis (TB; in partnership with the TB Alliance), Chagas disease and visceral leishmaniasis (in partnership with the Drugs for Neglected Diseases *initiative* (DNDi)).

Specific criteria have been proposed for defining hits and leads in the development of drugs for diseases such as malaria<sup>4-6</sup>; however, it is essential that such criteria are regularly reviewed and updated as part of evolving target product profiles (TPPs) to reflect accumulated experience from drug discovery projects and emerging scientific research, clinical experience, policy guidance and patient need. In order to reach a consensus on the HTL criteria to be used to guide its collaborative activities, the GHIT Fund recently convened an initial gathering of its key international partners involved in drug discovery — the MMV, the TB Alliance and the DNDi — together with representatives from the Bill & Melinda Gates Foundation. The objective of this meeting was to take the first steps towards creating a shared, flexible strategy to expedite the discovery of the next generation of drugs for these diseases by GHIT Fund collaborations. Here, after briefly discussing some of the general characteristics of the relevant screening assays and established target product profiles, we present the proposed generic and disease-specific criteria for hits and leads. We hope that these guidelines may also be valuable to drug discovery efforts outside these partnerships.

## Hit discovery and assay development

For infectious diseases, hit candidates usually come from screens that involve intact pathogens, akin to phenotypic screens (or, more generally, high-content screens) rather than target-based screens. Which of these approaches is more productive is still being debated<sup>7,8</sup>. Despite being fuelled by advances in genomics, target-based high-throughput screening, along with computer-assisted modelling, has not been as productive as phenotypic screening in the antibacterial area<sup>9</sup>. However, for subsequent optimization efforts there is no doubt that knowing the molecular target of a hit series is a major advantage. For instance, such knowledge allowed the clinical antimalarial candidate DSM265 to be specifically optimized to inhibit *Plasmodium falciparum* dihydroorotate dehydrogenase rather than its human orthologue<sup>10</sup>. Additionally, knowing the molecular target enables target-specific liabilities to be identified earlier in the R&D process. Furthermore, from a

## Box 1 | The Global Health Innovative Technology Fund: unlocking Japanese innovation

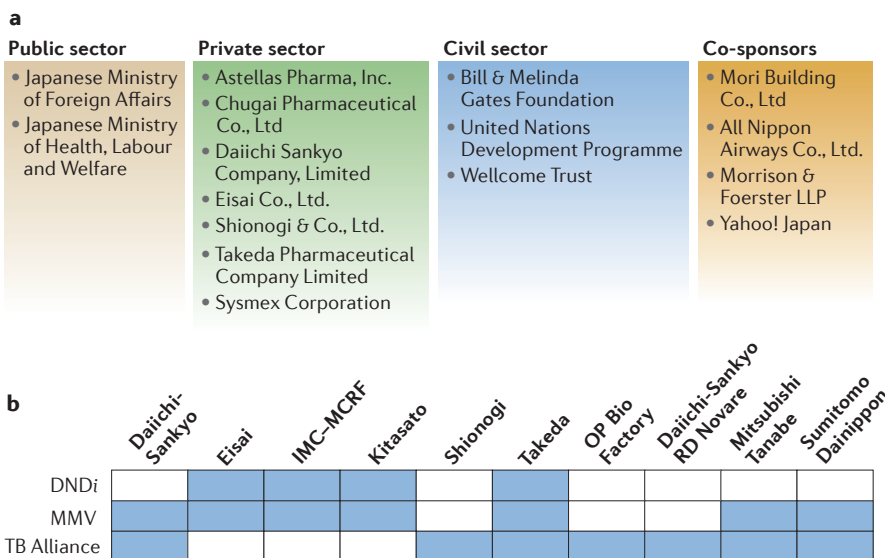
The Global Health Innovative Technology (GHIT) Fund of Japan was founded in April 2013 with a vision to alleviate the burden of infectious disease that prevents billions of people from seeking the level of prosperity and longevity commonly expected in industrialized nations. Uniquely, the GHIT Fund aims to do this by facilitating international partnerships that enable Japanese technology, innovations and insights to have a more direct role in making this vision a reality<sup>67,68</sup>. Along with six Japanese pharmaceutical companies and one diagnostic company, the GHIT Fund is supported by two Japanese government ministries and numerous unexpected co-sponsors, including Mori Building Co., Ltd., All Nippon Airways Co., Ltd., Morrison & Foerster LLP, and Yahoo! Japan (see the figure, part a). The Japanese pharmaceutical industry is represented by the 69 members of the Japanese Pharmaceutical Manufacturers Association.

A key aim of the GHIT Fund is to link existing product development partnerships (PDPs) — responsible for discovering and delivering new medicines for malaria, tuberculosis (TB) and neglected tropical diseases (NTDs) — with Japanese partners that have expertise in drug research and development (R&D). This is achieved by establishing a portfolio of screening projects and a hit-to-lead (HTL) platform that use defined criteria to ensure that the most attractive chemical series are selected for further optimization and development. Research collaborations must comprise at least two distinct organizations, one of which is Japanese, in order to be eligible. For the screening and HTL platforms the compounds must originate from a Japanese entity. In addition, the HTL platform aims to maximize the skills and resources of existing global PDPs that operate in the field of NTDs and other infectious diseases that affect people in the developing world, such that each new partner must enter into a research collaboration with one of three specific PDPs (the Medicines for Malaria Venture (MMV), the Drugs for Neglected Diseases initiative (DNDi) or the TB Alliance).

Beyond the benefits of sharing expertise, collaborations of this type will allow diverse chemical libraries to be probed using various screening approaches. Importantly, Japanese compound assets of synthetic and natural origin could be a rich source of novel and chemically diverse compounds that have not previously been screened for infectious diseases of the developing world. Past examples of key natural-product-based drugs that originated from Japanese discovery efforts include ivermectin, an antiparasitic drug from the actinomycete *Streptomyces avermectinus*<sup>69–71</sup>; the multiple sclerosis therapy fingolimod, which is a metabolite of the insect fungus *Isaria sinclairii*<sup>72</sup>; and the founder of the statin drug class, mevastatin, which was discovered in *Penicillium* spp. extrolites by Akira Endo of Sankyo in the 1970s<sup>73,74</sup>.

In just over 2 years since its inception, the GHIT Fund has facilitated more than 30 partnerships (for screening collaborations see the figure, part b) and invested more than US\$40 million in them. All proposals received are evaluated by external reviewers and a selection committee, both of which are completely independent from the GHIT Fund.

In addition to the existing three platforms (the screening, HTL and product development platforms), the GHIT Fund also initiated the 'Target Research Platform in Partnership with Grand Challenges' (TRP) for early-stage development of radically new and improved drugs, vaccines and diagnostics to prevent and treat infectious diseases that are prevalent in developing countries. With additional Japanese partnerships expected and additional compound screening and HTL proposals being submitted<sup>67,68</sup>, the GHIT Fund encourages research and shows a commitment to the cause that is unprecedented in this area of research within Asia.



IMC, Institute of Microbial Chemistry; MCRF, Microbial Chemistry Research Foundation.

portfolio point of view, it ensures that a mix of mechanistic approaches are followed, which is especially relevant in diseases in which drug resistance develops and spreads quickly (for example, TB and malaria).

Fortunately, advances in genomics and related technologies usually enable the molecular target of a drug (or at least its mode of action) to be elucidated even when the compound was discovered in a phenotypic screen; such screens may therefore be viewed as discovery engines for druggable targets, along with the drug molecules themselves. The two crucial success components for screening are the availability of chemical libraries with 'interesting' chemistry and taking great care to develop assays that faithfully reproduce the microenvironment of a pathogen with a disease-relevant readout, while also maintaining throughput and robustness. Such advanced assays also eliminate compounds that do not penetrate cellular membranes or are otherwise unavailable to the target.

The organisms that cause malaria, TB and NTDs are biologically quite diverse, including viruses, bacteria and eukaryotes. Each of these comes with its own challenges and opportunities, resulting in differences in hit and lead criteria and assay conditions. Dogma holds that bacteria replicate quickly (~20-minute generation time) and eukaryotic (mammalian) cells double in number in ~24 hours, but these guidelines do not apply to the pathogens that cause the four diseases discussed in this article. *Mycobacterium tuberculosis* is an exception among bacteria in that it has up to 16-hour doubling times, whereas the doubling times of the eukaryotic *Leishmania* spp. parasites, which cause visceral leishmaniasis, are ~6–9 hours. The *in vitro* culture of all major strains of *P. falciparum* — the species responsible for the majority of deaths from malaria — has been an important breakthrough<sup>11</sup>, but its doubling time is ~48 hours<sup>12</sup>. Owing to these differences, biomass generation and assay set-up vary in difficulty between organisms.

The potency cut-offs for compounds to progress to the next stage of development varies by disease (see below) and is decided by a number of factors. Among these factors are empirical hit rates (the higher the hit rate, the lower the cutoff concentration). *Plasmodium* spp. hit rates are generally good, which is possibly related to the unusual (and vital) apicoplast organelles that these organisms harbour. Moreover, these organisms must actively overcome the toxicity of haem degradation and remodel their host cell (red blood cell) membrane transport capabilities, and these processes are druggable. By contrast,

it is perhaps not surprising that targeting and killing the intracellular amastigote form of *Leishmania* spp. parasites is more challenging, as amastigotes reside inside acidic phagolysosomes within macrophages, which present additional membranes and pH gradients that drugs must cross before reaching their target. *Trypanosoma cruzi*, the parasite that causes Chagas disease, presents its own challenges owing to its ability to infect many cell types, producing a highly dynamic infection. Furthermore, drugs that only target the replicating stages of the parasite may leave non-replicating forms, such as trypomastigotes, capable of maintaining infections long after the end of the treatment<sup>13</sup>.

### Established target product profiles

It is important that hit and lead series are assessed as early as possible for conformity with the relevant disease TPPs and, if appropriate, target candidate profiles (TCPs). However, in most cases, further detailed biological or pharmacokinetic studies will be required to fully judge how a series can be strategically positioned. These TPPs and TCPs are typically developed by the PDPs in discussion with the research and medical communities.

**TPPs and TCPs for malaria.** In the case of malaria, two TPPs and five TCPs have been established, which reflect different patient populations and medicinal uses<sup>6,14</sup>. In brief, malaria TPP1 is focused on treatment, ideally from a single dose, such that all symptomatic and asymptomatic parasites (gametocytes) in a host are cleared and transmission is blocked in addition to the patient being cured. TPP2 is focused on prophylaxis, because when attempting to eliminate a disease it is recognized that people in once-endemic regions, now with potentially reduced immunity, may require protection in the event of transmission outbursts. The TPPs focus on the profile of a medicine that is composed of two or more active ingredients (a criterion that is presently mandatory for artemisinins<sup>15</sup>). By contrast, the TCPs focus on attributes that individual molecules need to possess (acknowledging that one molecule may possess more than one attribute). The key features of the five TCPs are: fast parasite clearance (TCP1); a combination partner, ideally long duration, which provides post-treatment prophylaxis (TCP2); prevention of *Plasmodium vivax* and *Plasmodium ovale* relapse (TCP3a); prevention of transmission (TCP3b); and chemoprotection (TCP4)<sup>6</sup>. TCP1, TCP2, TCP3a and TCP3b all support TPP1, whereas TCP4 supports TPP2 (REF. 6).

### Box 2 | Generic hit selection criteria for infectious diseases

The panel of drug experts convened by the Japanese Global Health Innovative Technology (GHIT) Fund identified the following generic hit selection criteria:

- A hit should have a potency that is consistent with the potential to deliver a lead compound (see BOXES 4–6 for details of criteria defined specifically for each disease)
- The chemical structure of a hit should be confirmed by identification (for natural products), re-synthesis or re-purification
- Primary screening data should be validated on a selection of hit compounds (>90% pure)
- A hit should have an acceptable *in vitro* response (typically, a sigmoidal concentration–growth inhibition curve reaching a maximal 100% efficacy, with a Hill coefficient ideally between 0.5 and 1.8)
- Preliminary knowledge of the structure–activity relationship (SAR; often available from analogues in the original screening library) of a hit is desirable
- A hit should have a tractable chemotype: it should have no highly reactive or unstable moieties in the pharmacophore and be amenable to structural variation by chemical (or biochemical) synthesis. Hits should pass basic drug-like filters, such as pan-assay interference filters (PAINS)<sup>75</sup>, to eliminate promiscuous hits that lack target specificity<sup>76,77</sup>. Conformity to the ‘rule of five’ (REF. 58) is preferred
- There should be a greater than 10-fold selectivity window for cytotoxicity using a mammalian cell line (for example, HepG2 or Vero cells)
- A hit requires adequate selectivity in a biochemical counter-assay (for example, a homologous mammalian target) where relevant. However, most infectious disease hit-to-lead projects are not target-based screens but phenotypic
- No serious intellectual property conflicts should exist (that is, a ‘freedom to operate’ is needed). However, with the value of US Food and Drug Administration (FDA) priority review vouchers now entering the pharmacoeconomic equation, there are further possibilities to develop drugs for infectious diseases even in absence of intellectual property protection
- No major synthesis or formulation issues should be anticipated (compounds should ideally be synthesized in ≤5 steps with an acceptable yield and acceptable solubility). For reasons of affordability, this criterion is more stringent than for drug discovery in general

**TPPs for TB.** TPPs for TB usually require new drugs to shorten the duration of treatment, demonstrate efficacy against drug-sensitive and drug-resistant *M. tuberculosis* strains and show potential for use in drug combinations in developing countries (for example, those with oral and once-daily dosing, and low cost of goods). These goals are based on the fact that the majority of TB drugs and treatments have not changed in the past half-century, even though two new drugs (bedaquiline<sup>16</sup> and delamanid<sup>17</sup>) were recently approved — the first new drugs for TB in the past 40 years. Additional drugs are needed to develop new treatment regimens.

**TPPs for Chagas disease and leishmaniasis.** The DNDi has defined a TPP for Chagas disease with ‘acceptable’ and ‘ideal’ criteria (see the [DNDi website](#)). The benchmark for clinical efficacy is benznidazole, and this should therefore be the comparator for all hit and lead candidates. Similarly, the DNDi has published (with consultancy) ‘optimal’ and ‘minimal’ TPPs for visceral and cutaneous leishmaniasis. For visceral leishmaniasis, a safe oral drug with >90% efficacy within 10 days is crucial. Relevant for early compound triage is that a drug must be active against all resistant strains.

### Generic criteria for hits and leads

**Hit series definition and criteria.** A number of generic hit criteria, identified by a panel of drug experts convened by the GHIT Fund, apply to all four infectious diseases, and these are listed in BOX 2. In general, the objectives at this stage are to build confidence in the quality of a compound series and the associated data, understand the liabilities and, if possible, generate data that guide medicinal chemists during the HTL phase<sup>18</sup>. Two main types of criteria are covered: disease-specific criteria, focusing on potency, efficacy and pathogenicity; and compound-specific criteria, which focus principally on the chemical scope of the compound and a risk assessment of drug metabolism and pharmacokinetics (DMPK), as well as the physical properties that are predictive of an oral therapy.

**Lead series definition and criteria.** During the HTL phase, the project team focuses on the optimization of a chemical series so as to improve any compound properties that could be an obstacle to further progression. Depending on the profile, the HTL strategy and medicinal chemistry plan can be very different between series. The milestone for



## Box 3 | Generic lead selection criteria for infectious diseases

The panel of drug experts convened by the Japanese Global Health Innovative Technology (GHIT) Fund identified the following generic lead selection criteria:

- A lead requires an acceptable *in vitro* potency for the relevant disease (see BOXES 4–6 for disease-specific details). In general, high potency (a low half-maximal inhibitory concentration ( $IC_{50}$ )) is highly desirable but not at the expense of poor physicochemical properties or drug metabolism and pharmacokinetic (DMPK) characteristics
- A lead should have oral efficacy in the appropriate disease model. Oral efficacy removes one key risk early on, namely, uncertainty about *in vivo* validation of the mechanism of action of the series (which is often unknown at this stage)
- The synthetic chemistry should be amenable to series expansion, as many more compounds are likely to be needed for testing in as short a time frame as possible
- A lead needs a greater than 10-fold selectivity in killing pathogens as opposed to mammalian cells in tuberculosis (TB) and a greater than 100-fold selectivity for the pathogens that cause the other diseases (malaria, Chagas disease and visceral leishmaniasis). This reflects the increased challenge of finding potent, attractive compounds that target TB
- A lead should have acceptable physicochemical properties (typically, solubility in phosphate-buffered saline  $>10\mu\text{M}$ ; a sufficient level of solubility is expected to avoid problematic formulations<sup>78</sup>. Acceptable lipophilicity; LogP values are typically  $<5$  and ideally  $<3$ )<sup>58</sup>
- A lead requires manageable drug metabolism and pharmacokinetic profiles. This involves liver microsomal and hepatocyte stability across species; understanding of the plasma protein, microsomal and media binding; good membrane permeability; and no unmanageable cytochrome P450 inhibition
- A lead needs to demonstrate good oral bioavailability in rodents (demonstrated  $F>25\%$ ); ultimately, high oral bioavailability is highly desirable as it will reduce the potential for inter-patient variability, pill size, dose and cost of the medicine. It is therefore crucial that this parameter is tractable at the lead stage
- A lead needs an acceptable early safety assessment based on target (orthologue) and compound liabilities, *in vivo* observations, *in vitro* studies (for example, genotoxicity and the mini-Ames test), cytotoxicity, cardiac safety (as assessed using the hERG channel (QT prolongation)) and *in silico* approaches. A secondary pharmacology selectivity profile, consistent with achieving selectivity with the candidate compound, is also required. This would include human orthologues and paralogues of the targeted enzyme or receptor, if known; the number of these that are to be tested depends on the gene family size and their known or suspected safety risks (when targeted)
- All liabilities of the series should be understood (as a result of extensive profiling) and a rationale and medicinal chemistry plan generated for how they might be overcome in the subsequent optimization phase
- A lead should contain no known toxicophores or undesirable reactive groups and no detrimental chemical feature or characteristic associated with the pharmacophore indicative of, for example, adduct formation. This avoids the scenario in which a region of the molecule responsible for activity has a liability that cannot be overcome
- A lead should display no acute toxicity in *in vivo* efficacy studies. Although no formal *in vivo* safety studies are performed at this stage, careful observation of efficacy studies, particularly at high, repeated doses, can be informative
- Preferably, there should be no apparent intellectual property obstacles to the progression of any series (freedom to operate)

this phase is the delivery of a lead series. The generic criteria for such a lead are shown in BOX 3 (REF. 19).

### Disease-specific hit and lead criteria

Guided by the specific requirements for each disease as well as existing target product and candidate profiles, the committee coordinated by the GHIT Fund devised disease-specific criteria for hits and leads, which are discussed below and summarized in BOXES 4–6.

**Malaria.** There are several good treatments for malaria, but the challenge of emerging drug resistance is ever present, particularly

in South-East Asia where cases of increased parasite-clearance times are on the rise in patients treated with combination therapies that include artemisinin derivatives (for example, artesunate). This has revealed a new type of *Plasmodium* spp. resistance — essentially one in which the ring stage of the intra-erythrocytic cycle can tolerate drug intervention<sup>20–23</sup>. As the mainstay malaria treatments rely on a component that is artemisinin-based, a global public health catastrophe could emerge unless new drugs that overcome this risk are delivered.

In addition, current antimalarials generally target the asexual blood stage of the disease.

Breaking the life cycle and killing parasites during other asymptomatic phases — such as the sexual-stage gametocytes and liver-stage hypnozoites and schizonts — will be crucial for providing treatments to block relapse in *P. vivax* infections, preventing transmission and protecting vulnerable populations. Indeed, only with such drugs can the global goal of eradicating malaria be realized. Extensive research on these two disease-causing pathogens (*P. falciparum* and *P. vivax*) over the past decade has helped to establish some very specific entry criteria for drug discovery programmes<sup>14</sup>.

The cellular potency criterion for a hit for malaria research (BOX 4) is based on the extensive screening efforts of the MMV and its partners over the past 7 years<sup>24</sup>. Through partnerships in both industry and academia, more than 5 million compounds have been screened against the asexual blood stages of *P. falciparum* and the cutoff potency from these *in vitro* screens has mostly been around 1–2  $\mu\text{M}$ . Nevertheless, more than 25,000 compound hits were available for follow-up<sup>25–27</sup>. Screening against liver stages of *Plasmodium* spp. has also delivered many potent hits<sup>28</sup>. Naturally, activity should be confirmed with a pure compound, and hits should have selectivity for the parasite over a mammalian cell line (for example, a greater than 10-fold difference between the  $IC_{50}$  and  $CC_{50}$ , the half-maximal inhibitory concentration against the parasites and the half-maximal cytotoxic concentration against the host mammalian cells, respectively) and display an acceptable concentration response, all of which provide confidence in a specific interaction and effect.

By the start of lead optimization there is a need for clarity on the TPPs and TCPs of the series, as these define the goal and tactics for the subsequent optimization phase. The potency required for progression depends on the TCP. For blood stages (TCP1 and TCP2), the crucial aspect is *in vivo* oral efficacy in the *P. falciparum* SCID (severe combined immunodeficiency) mouse model of infection<sup>29</sup>, with the key feature for TCP1 compounds being rapid parasite clearance *in vivo*, at rates at least as good as those of chloroquine. For prophylaxis (TCP4), the crucial aspect is *in vivo* oral efficacy in a *Plasmodium berghei* sporozoite challenge model (or equivalent)<sup>30</sup>. From experience, achieving a dose that eradicates 90% of the target pathogen ( $ED_{90}$ )  $<50\text{ mg per kg}$  provides confidence that the series has a parasitological foundation for the lead optimization phase; ultimately, a successful candidate will have an  $ED_{90}$   $<10\text{ mg per kg}$  (often considerably lower). Typically,

this is achievable if a compound has an  $IC_{50}$  <100 nM against the blood or liver stages and appropriate DMPK properties. However, a compound may still be acceptable for follow up if one of these components does not meet the proposed criteria, provided the other parameter is outstanding.

Given the risk of resistance and a need for drugs with novel modes of action, potency *in vitro* across a panel of established drug-resistant parasite strains isolated from patients is also crucial. The genetics underlying the emergent resistance against artemisinins, as seen in the Mekong area in South-East Asia, are becoming increasingly understood<sup>31–33</sup>, and this knowledge is being used to set up panels of parasites against which new drug candidates can be tested. For TCP1 at least, new drugs are most likely to be used in combinations (as presently mandated for all artemisinins). This means that a newcomer must be evaluated for potential drug–drug interactions with other TCP1 antimalarials.

For TCP3a (anti-relapse), it is necessary to have *in vitro* data supporting the activity of compounds on hypnozoites<sup>34</sup>, whereas for TCP3b additional activity against mature and, ideally, early-stage gametocytes is required in addition to asexual blood-stage potency<sup>35</sup>.

**Tuberculosis.** *M. tuberculosis* is so well adapted to its human host that almost one third of the world population is estimated to be infected. Although only 10% of infected individuals are believed to develop TB in their lifetime, this still results in ~1.5 million deaths annually<sup>36</sup> (see the [World Health Organization \(WHO\) website](#)). Therefore, there is an urgent need for new anti-TB drugs.

It still takes 6 months to cure drug-sensitive TB (DS-TB) and a minimum of 18 months to treat multidrug-resistant TB (MDR-TB). It is highly desirable to shorten treatment duration for both patients with DS-TB and those with MDR-TB to improve compliance and to limit the spread of drug-resistant TB (DR-TB). Standard care of patients with DS-TB includes a combination of isoniazid, rifampicin, pyrazinamide and ethambutol for the first 2 months and a combination of isoniazid and rifampicin for the remaining 4 months. Patients with MDR-TB, whose *M. tuberculosis* strains are resistant to isoniazid and rifampicin, are treated with second-line TB drugs that include aminoglycosides, quinolone antibiotics, cycloserine and capreomycin. An updated dataset was recently published for the most commonly used TB drugs with respect to *in vitro* potency, cidalty, physicochemical and

#### Box 4 | Summary of main checkpoint criteria for antimalarial hits and leads

Guided by the specific requirements for the disease, and taking into account existing target product and candidate profiles, the committee coordinated by the Global Health Innovative Technology (GHIT) Fund devised the following disease-specific criteria for hits and leads for malaria.

##### Validated hit

- Cellular potency criteria: hits should have an effector concentration for half-maximum response ( $EC_{50}$ ) <1  $\mu$ M for sensitive and multiple resistant strains of *Plasmodium* spp.
- Cytotoxicity criteria: hits require a greater than 10-fold selectivity between the half-maximal cytotoxic concentration ( $CC_{50}$ ) for the mammalian cell line and the  $EC_{50}$  for *Plasmodium* spp.

(See also BOX 2 for the evaluation criteria regarding the acceptability of chemical structure, novelty and confirmation)

##### Early lead

- Cellular potency criteria: a lead requires  $EC_{50}$  <100 nM for sensitive and multidrug-resistant strains of *Plasmodium* spp.
- Cytotoxicity criteria: a lead should have a greater than 100-fold selectivity between mammalian cell line  $CC_{50}$  and *Plasmodium*  $EC_{50}$ . Frontrunners should be tested across the malaria life cycle and key mechanistic assays so as to ensure an understanding of the phenotype and target candidate profile (TCP) potential of each series and (preferably) novel mechanisms of action
- *In vivo* efficacy criteria: when administered orally in the blood stages of infection (TCP1 and TCP2), a lead should achieve parasite clearance at a dose that eradicates 90% of the target pathogen ( $ED_{90}$ ) <50 mg per kg (typically four doses over 4 days) in the *Plasmodium falciparum*-infected SCID (severe combined immunodeficiency) mouse model. TCP1 should demonstrate a rapid rate of parasite clearance. For TCP3a, the anti-relapse TCP, there are no *in vivo* criteria for leads (in the absence of good models), but a lead should demonstrate anti-hypnozoite activity *in vitro*. For TCP3b, the transmission-blocking TCP, a lead should demonstrate potency in a gametocyte assay (for example, as measured by gamete formation<sup>79</sup>) in a similar range to that of the *in vitro* asexual blood stage potency. For TCP4, the chemoprotection TCP, a lead should have efficacy in a prophylaxis model of malaria, with  $ED_{90}$  <50 mg per kg

pharmacokinetic properties generated under standardized conditions<sup>9</sup>. For patients treated for MDR-TB, the cure rate is only ~48% and this needs to be improved drastically, whereas the cure rate of patients treated for DS-TB is ~85%<sup>9</sup>.

There has been a steady increase in TB drug resistance; the WHO estimated that 3.5% of new cases and 21% of previously treated cases are MDR-TB<sup>36</sup>. Among patients with MDR-TB, an estimated 9% have extensively drug-resistant TB (XDR-TB), meaning that the pathogens are resistant to all second-line TB drugs<sup>36</sup> in addition to rifampicin and isoniazid. A new drug must show efficacy against all resistant strains, which calls for agents with novel mechanisms.

The new candidates must also be orally available to ensure wide usage, especially in developing countries. For the same reason, the cost of goods needs to be low and pill size (which is determined by potency and pharmacokinetic properties) reasonable. Furthermore, as they will be used in combination with other TB drugs to stem the emergence of resistance, the new drugs must have a low risk for drug–drug interactions. As patients with TB are often co-infected with HIV, new agents also need to be compatible with most of the anti-retroviral

drugs. All of these requirements are especially stringent because TB drugs are administered for extended time spans.

Treatment shortening can only be achieved with an agent or a regimen that effectively eliminates non-replicating *M. tuberculosis*<sup>9</sup>. Non-replicating *M. tuberculosis* is metabolically less active and less susceptible to antibiotics than actively replicating bacteria<sup>9</sup>, and new hits must be evaluated for their capacity to kill both replicating and non-replicating *M. tuberculosis*, as measured in the microplate Alamar Blue assay (MABA<sup>37</sup>) and low oxygen recovery assay (LORA<sup>38</sup>), which are both suited for high-throughput screening. The desired hit and lead profiles are summarized in BOX 5.

The most-recent TB hits were discovered in phenotypic screens with cultured *M. tuberculosis*, typically using the MABA<sup>9</sup>. In addition, low-oxygen culture conditions (LORA<sup>38</sup>) and other assays<sup>39,40</sup> are used to identify hits that kill slow- or non-replicating bacteria<sup>41</sup>. However, these assays do not fully capture the little-understood mechanisms whereby TB bacteria move in and out of latency<sup>42–44</sup>. In addressing this need, a new assay was recently developed for agents that kill *M. tuberculosis* bacteria that reside in macrophages<sup>45</sup>, resulting in the discovery of Q208 (REF. 46).

**Box 5 | Main checkpoint criteria for anti-tuberculosis hits and leads**

Guided by the specific requirements for the disease, and taking into account existing target product and candidate profiles, the committee coordinated by the Global Health Innovative Technology (GHIT) Fund devised the following disease-specific criteria for hits and leads for tuberculosis (TB).

**Validated hit**

- Cellular potency criteria: a hit should have a minimum inhibitory concentration (MIC) against the *Mycobacterium tuberculosis* laboratory strain H37Rv of  $<10\ \mu\text{M}$  under replicating growth conditions, for example, under microplate Alamar Blue assay (MABA) conditions<sup>37</sup>
- A hit requires a greater than a 10-fold difference between cytotoxicity against mammalian cells (Vero cells) and the MIC against *M. tuberculosis* H37Rv
- Hits should show evidence of a preliminary structure–activity relationship (SAR) among analogues (See also BOX 2 for the evaluation criteria regarding the acceptability of chemical structure, novelty and confirmation)

**Early lead**

- *In vitro* potency criteria: a lead should have a MIC against the *M. tuberculosis* strain H37Rv of  $<1\ \mu\text{M}$  under MABA conditions and a MIC  $<10\ \mu\text{M}$  under non-replicating anaerobic conditions (low-oxygen-recovery assay, LORA<sup>38</sup> conditions)
- A lead should exhibit *in vitro* activity against *M. tuberculosis* strains that are resistant to a single TB drug, such as isoniazid or rifampicin, indicating a new mechanism of action
- A demonstration of the *in vitro* bactericidal activity of a lead is required as indicated by time–kill curves showing that the number of colony-forming units (CFUs) decreases over time (for example, 14 days) compared with the number of CFUs at the beginning of the experiment
- The oral bioavailability of a lead must be demonstrated in rodents
- A demonstration of the oral efficacy of a lead in a mouse acute infection model is required<sup>80</sup>
- A preliminary indication of the safety of a lead must be demonstrated (in hERG<sup>81</sup> and cell health assays<sup>82</sup>)

The second challenge in discovering drugs that target latent TB requires an understanding of TB pathogenesis. *M. tuberculosis* primarily infects and replicates in activated macrophages but can persist in a non-replicating state in foamy macrophages<sup>47</sup>. This host–pathogen system can form granulomas in which infected macrophages are surrounded by layers of active macrophages and other immune cells. The central granuloma region is oxygen-deprived and necrotized but contains latent *M. tuberculosis*; in the lung this results in cavities. The inner necrotic area is poorly irrigated, complicating local drug exposure. Granulomas do not normally form in *in vivo* models; however, recently, animal models of granulomas were developed<sup>48,49</sup>. In one of these models, it was shown that treatment with a vascular endothelial growth factor (VEGF)-specific antibody can restore vasculature and drug access in TB granulomas<sup>49</sup>. Continued research into new assays and models is crucial, as well as pursuing completely different approaches that aim to stimulate host immune responses against TB<sup>50</sup>.

**Visceral leishmaniasis and Chagas disease.**

For visceral leishmaniasis, existing drugs have variable efficacy and serious toxicities<sup>51</sup>; only one (miltefosine) is administered orally and the others are given by intravenous or intramuscular injections, which are impractical.

The goal for research must be to transform patient therapy from poorly adapted antimonial treatments (for example, sodium stibogluconate (SSG)) to simple, patient-adapted oral therapies that are affordable, safe and efficacious in both children and adults.

In the past 15 years, combinations of drugs with similar or improved efficacy to the older antimonials, but with improved safety and tolerability profiles, have been developed. These include combinations containing liposomal amphotericin B, paromomycin and/or miltefosine<sup>52</sup>. However, these drugs remain costly, are difficult to administer, have poor stability at the high temperatures that occur in endemic regions, require lengthy treatment and/or are poorly tolerated. In addition, there is a dichotomy in drug efficacy in regions of the world where visceral leishmaniasis is endemic. In South Asia, the medical needs for visceral leishmaniasis are presently moderately to well met. However, in East Africa and Latin America, the efficacy and tolerability of current visceral leishmaniasis therapies remain a challenging area for improvement.

Ideally, what is needed in the treatment of visceral leishmaniasis is a simple oral combination therapy that would prove to be advantageous and/or effective in maintaining or improving efficacy, improving tolerability and preventing or delaying the emergence of

resistance. Furthermore, a treatment adapted to field conditions, with a shorter treatment duration and that could be used globally, would be optimal. In addition, to address the development of resistance by the parasite to drug monotherapies — as documented for current therapies — drugs with new and distinct mechanisms of action should be developed as combination therapies early in development and not used as monotherapies.

In the case of Chagas disease, there are even fewer treatment options than there are for visceral leishmaniasis. Monotherapy with nifurtimox or benznidazole (both from the same nitroheterocycle class) remains the only recognized treatment, but these agents require long treatment courses, have variable efficacy and cause serious side effects, resulting in discontinuation of treatment in ~20–30% of patients. It is crucial that new classes of effective, well-tolerated, orally acting and short-course treatments are progressed into the clinic to provide improved options for patients. New classes of drugs for Chagas disease are also essential to enable the development of combination therapies to improve efficacy and toleration, reduce treatment duration and combat the risk of the development of resistance to monotherapies. Recently, a trial in chronic Chagas disease was described for posaconazole, an inhibitor of *T. cruzi* 14- $\alpha$  demethylase (CYP51) that has a different mechanism of action to the currently used nitroheterocyclics, as exemplified by benznidazole<sup>53</sup>. In spite of good preclinical efficacy, the new compound did not meet expectations, resulting in more treatment failures than benznidazole. It is believed that the standard Chagas mouse model that was used to validate posaconazole only represents the early acute phase of infection; thus, it is important to use models and assays that also allow testing of the chronic late stage of the disease<sup>13,54</sup>.

In order to promote the discovery of safe, efficacious and orally acting treatments for visceral leishmaniasis and Chagas disease that overcome the limitations of existing regimens, discovery needs to focus on new chemical series, leaving behind the flawed classes in current use. Until recently, the steady identification of new chemical series has been hindered by a limited screening capacity coupled with very low hit rates. However, progress has been made and larger compound collections are now being screened. More than a million compounds have been screened against *T. cruzi* and *Leishmania* spp.<sup>54–56</sup>, but the rate of hit discovery still lags far behind that for malaria. It remains necessary to avoid discarding precious chemical series by setting the hit



selection criteria too high. Thus, hit criteria focus on the identification of new series and mechanisms of action with even modest activity against the intracellular forms of *Leishmania donovani* and *T. cruzi*, the causative agents of leishmaniasis and Chagas disease, respectively<sup>57</sup>.

Although this inclusive approach to hit selection provides scope for subsequent HTL projects it does bring with it a high rate of early attrition and often leads to long optimization campaigns before quality leads are produced. The HTL entry criteria stated in BOX 6 are a combination of the current well-documented 'best practices' for targeted drug discovery projects<sup>58</sup> and a modest phenotypic *in vitro* activity hurdle<sup>59</sup>. Once the encouraging *in vitro* anti-parasitic activity of a new series can be coupled with sufficient *in vivo* plasma exposure, ideally following oral dosing, the next hurdle is to demonstrate a robust reduction in the parasite burden at target organs in infected rodents, which is the defining characteristic of a lead that is ready for subsequent optimization. Unfortunately, we have little knowledge regarding how *in vitro* potency correlates with *in vivo* activity in visceral leishmaniasis and Chagas disease, with a lack of pharmacokinetic and pharmacodynamic (PK/PD) examples. For *T. cruzi* infection, oral dosing is required for *in vivo* proof-of-concept studies, which is in line with the requirement for orally acting therapies. Although oral dosing is also preferred for studies with rodents infected with *Leishmania* spp., intraperitoneal or intravenous administrations are also acceptable given that a short-course treatment administered by the parenteral route could fulfil the TPP for visceral leishmaniasis.

## Discussion

Our consultation among drug discovery experts collaborating with the GHIT Fund revealed differences in the criteria considered most conducive to cost-efficient drug discovery, which are associated with the specific requirements for the diseases reviewed here. One of the dangers in establishing stringent, clear-cut criteria for entry into the HTL phase and then progression into lead optimization is that valuable chemicals are discarded at an early stage. It is obvious that a molecule such as artemisinin, were it to emerge today as a hit in a high-throughput screening campaign, would severely struggle to be taken forward, scoring extremely low in chemical tractability, ease of synthesis and chemical suitability (the compound has a highly reactive peroxide). Even if it were taken further (without analysis of the

### Box 6 | Hit and lead criteria for Chagas disease and visceral leishmaniasis

Guided by the specific requirements for the disease, and taking into account existing target product and candidate profiles, the committee coordinated by the Global Health Innovative Technology (GHIT) Fund devised the following disease-specific criteria for hits and leads for Chagas disease and visceral leishmaniasis.

#### Hit and lead criteria for Chagas disease

- Key hit selection criteria: cellular potency — a hit should have a half-maximal inhibitory concentration ( $IC_{50}$ ) <10  $\mu$ M against intracellular *Trypanosoma cruzi*
- Key lead selection criteria: in an acute mouse model of Chagas disease, a hit should demonstrate an 80% reduction of parasite burden in organs or tissues, or there should be no parasites detected at the end of treatment and an increase in lifespan with up to 10 doses at 50 mg per kg delivered orally

#### Hit and lead criteria for visceral leishmaniasis

- Key hit selection criteria: cellular potency — a lead should demonstrate an  $IC_{50}$  <10  $\mu$ M against intracellular *Leishmania donovani*
- Key lead selection criteria: in a mouse (or hamster) model (infected with *L. donovani* or *Leishmania infantum*), treatment with a lead should result in a >70% reduction in liver parasite burden after at most 5 doses at 50 mg per kg delivered orally once or twice per day

structure–activity relationship (SAR)), scientists would discover that the compound has a very short *in vivo* half-life (minutes), is metabolically unstable and its mode of action and target are poorly understood<sup>60</sup>. Along with serious preclinical safety concerns<sup>61,62</sup>, there is every reason to believe that the modern drug discovery process would have discarded the drug long before it had a chance to save the lives of millions. In fact, algorithms that select 'drug-like' molecules generally exclude artemisinin-like molecules from the chemical libraries used for HTS in the first place. Chemical tractability is important, but it must be kept in mind that there are many other molecules that have progressed straight from a phenotypic screen (that is, without further chemical modification) into patients, such as tamatinib (also known as R406)<sup>63</sup>, paclitaxel<sup>64</sup>, rapamycin<sup>65</sup> and cyclosporine<sup>66</sup>.

Conversely, there are numerous examples in which chemical series were pursued far too long (in hindsight), soaking up valuable resources, time and careers that should have been invested elsewhere. As an additional complication in navigating between these two extremes, the TPPs may shift with changes in the clinical landscape for infectious diseases that affect people in the developing world; for example, such changes may include a shift in policy from curing patients ad hoc to eradication programmes that require mass drug administration, a decision to focus on the prevention of transmission rather than on cures, changing disease priorities and pharmacoeconomics, the emergence of co-infections (with concomitant drug–drug interaction risks) or the spread of resistance.

Nevertheless, in the midst of all these complexities, project teams can attempt to steer their compounds using the set of criteria and

considerations presented here in the light of the four sets of disease TPPs as beacons. Our analysis has split recommendations for entry into HTL and lead optimization phase into a generic part and a more specific part for each disease. As stated earlier, these guidelines were established in the context of GHIT Fund-coordinated collaborations, but we hope that they will find wider adoption and aid the acceleration and expansion of pipeline drug candidates for these serious diseases.

Kei Katsuno and B. T. Slingsby are at the Global Health Innovative Technology (GHIT) Fund, Ark Hills, Sengokuyama Mori Tower (25F), 1-9-10 Roppongi, Minato-ku, Tokyo 106-0032, Japan.

Jeremy N. Burrows, Rob Hooft van Huijsduijnen and Dennis Schmatz are at the Medicines for Malaria Venture (MMV), 20, Route de Pré-Bois, 1215 Geneva 15, Switzerland

Ken Duncan and Peter Warner are at the Bill & Melinda Gates Foundation, PO Box 23350, Seattle, Washington 98102, USA.

Takushi Kaneko is at the Global Alliance for TB Drug Development (TB Alliance), 40 Wall Street, 24th Floor, New York, New York 10005, USA.

Kiyoshi Kita is at the University of Tokyo, 7-3-1, Hongo, Bunkyo, Tokyo 113-0033, Japan.

Charles E. Mowbray is at the Drugs for Neglected Diseases initiative (DNDi), 15 Chemin Louis-Dunant, 1202 Geneva, Switzerland.

Correspondence to K.K.  
e-mail: [kei.katsuno@ghitfund.org](mailto:kei.katsuno@ghitfund.org)

doi:10.1038/nrd4683

Published online 5 October 2015

1. World Health Organization. *Antimicrobial resistance: global report on surveillance 2014* (WHO, 2014).
2. Martis, E. A., Radhakrishnan, R. & Badve, R. R. High-throughput screening: the hits and leads of drug discovery — an overview. *J. Appl. Pharma. Sci.* **1**, 2–10 (2011).
3. Khanna, I. Drug discovery in pharmaceutical industry: productivity challenges and trends. *Drug Discov. Today* **17**, 1088–1102 (2012).
4. Nwaka, S. & Hudson, A. Innovative lead discovery strategies for tropical diseases. *Nat. Rev. Drug Discov.* **5**, 941–955 (2006).

5. Nwaka, S. *et al.* Advancing drug innovation for neglected diseases—criteria for lead progression. *PLoS Negl Trop Dis* **3**, e440 (2009).
6. Burrows, J. N., Hooft van Huijsduijnen, R., Möhrle, J. J., Oeuvray, C. & Wells, T. N. C. Designing the next generation of medicines for malaria control and eradication. *Malar J* **12**, 187 (2013).
7. Swinney, D. C. Phenotypic versus target-based drug discovery for first-in-class medicines. *Clin. Pharmacol. Ther.* **93**, 299–301 (2013).
8. Eder, J., Sedrani, R. & Wiesmann, C. The discovery of first-in-class drugs: origins and evolution. *Nat. Rev. Drug Discov.* **13**, 577–587 (2014).
9. Payne, D. J., Gwynn, M. N., Holmes, D. J. & Pompliano, D. L. Drugs for bad bugs: confronting the challenges of antibacterial discovery. *Nat. Rev. Drug Discov.* **6**, 29–40 (2007).
10. Deng, X. *et al.* Fluorine modulates species selectivity in the triazopyrimidine class of *Plasmodium falciparum* dihydroorotate dehydrogenase inhibitors. *J. Med. Chem.* **57**, 5381–5394 (2014).
11. Schuster, F. L. Cultivation of *Plasmodium* spp. *Clin. Microbiol. Rev.* **15**, 355–364 (2002).
12. Mutai, B. K. & Waitumbi, J. N. Apoptosis stalks *Plasmodium falciparum* maintained in continuous culture condition. *Malar J* **9** (Suppl. 3), S6 (2010).
13. Chatelain, E. Chagas disease drug discovery: toward a new era. *J. Biomol. Screen* **20**, 22–35 (2015).
14. Wells, T. N. C., Hooft van Huijsduijnen, R. & Van Voorhis, W. C. Malaria medicines: a glass half full? *Nat. Rev. Drug Discov.* **14**, 1–18 (2015).
15. Bosman, A. & Mendis, K. N. A major transition in malaria treatment: the adoption and deployment of artemisinin-based combination therapies. *Am. J. Trop. Med. Hyg.* **77**, 193–197 (2007).
16. Leibert, E., Danckers, M. & Rom, W. N. New drugs to treat multidrug-resistant tuberculosis: the case for bedaquiline. *Ther. Clin. Risk Manag.* **10**, 597–602 (2014).
17. Xavier, A. S. & Lakshmanan, M. Delamanid: a new armor in combating drug-resistant tuberculosis. *J. Pharmacol. Pharmacother.* **5**, 222–224 (2014).
18. Wunberg, T. *et al.* Improving the hit-to-lead process: data-driven assessment of drug-like and lead-like screening hits. *Drug Discov. Today* **11**, 175–180 (2006).
19. Hughes, M. *et al.* Early drug discovery and development guidelines: for academic researchers, collaborators, and start-up companies. *Assay Guidance Manual* [online], <http://www.ncbi.nlm.nih.gov/books/NBK92015/> (2012).
20. Arie, F. *et al.* A molecular marker of artemisinin-resistant *Plasmodium falciparum* malaria. *Nature* **505**, 50–55 (2014).
21. Ashley, E. A. *et al.* Spread of artemisinin resistance in *Plasmodium falciparum* malaria. *N. Engl. J. Med.* **371**, 411–423 (2014).
22. Burrows, J. Microbiology: malaria runs rings round artemisinin. *Nature* **520**, 628–630 (2015).
23. Mbengue, A. *et al.* A molecular mechanism of artemisinin resistance in *Plasmodium falciparum* malaria. *Nature* **520**, 683–687 (2015).
24. Burrows, J. N. & Waterson, D. in *Third World Diseases* (ed. Elliot, R. L.) 125–180 (Springer, 2011).
25. Gamo, F. J. *et al.* Thousands of chemical starting points for antimalarial lead identification. *Nature* **465**, 305–310 (2010).
26. Guigumde, W. A. *et al.* Chemical genetics of *Plasmodium falciparum*. *Nature* **465**, 311–315 (2010).
27. Plouffe, D. *et al.* In silico activity profiling reveals the mechanism of action of antimalarials discovered in a high-throughput screen. *Proc. Natl Acad. Sci. USA* **105**, 9059–9064 (2008).
28. Meister, S. *et al.* Imaging of *Plasmodium* liver stages to drive next-generation antimalarial drug discovery. *Science* **334**, 1372–1377 (2011).
29. Angulo-Barturen, I. *et al.* A murine model of falciparum-malaria by *in vivo* selection of competent strains in non-myelodepleted mice engrafted with human erythrocytes. *PLoS ONE* **3**, e2252 (2008).
30. Nilsen, A. *et al.* Quinolone-3-diarylethers: a new class of antimalarial drug. *Sci. Transl. Med.* **5**, 177ra37 (2013).
31. Winzler, E. A. & Manary, M. J. Drug resistance genomics of the antimalarial drug artemisinin. *Genome Biol.* **15**, 544 (2014).
32. Isozumi, R. *et al.* Novel mutations in K13 propeller gene of artemisinin-resistant *Plasmodium falciparum*. *Emerg. Infect. Dis.* **21**, 490–492 (2015).
33. Tun, K. M. *et al.* Spread of artemisinin-resistant *Plasmodium falciparum* in Myanmar: a cross-sectional survey of the K13 molecular marker. *Lancet Infect. Dis.* **15**, 415–421 (2015).
34. Demebele, L. *et al.* Towards an *in vitro* model of *Plasmodium* hypnozoites suitable for drug discovery. *PLoS ONE* **6**, e18162 (2011).
35. Sinden, R. E., Carter, R., Drakeley, C. & Leroy, D. The biology of sexual development of *Plasmodium*: the design and implementation of transmission-blocking strategies. *Malar J.* **11**, 70 (2012).
36. World Health Organization. Global tuberculosis report 2014. WHO [online], [http://www.who.int/tb/Publications/Global\\_Report/en/](http://www.who.int/tb/Publications/Global_Report/en/) (2014).
37. Franzblau, S. G. *et al.* Rapid, low-technology MIC determination with clinical *Mycobacterium tuberculosis* isolates by using the microplate Alamar Blue assay. *J. Clin. Microbiol.* **36**, 362–366 (1998).
38. Cho, S. H. *et al.* Low-oxygen-recovery assay for high-throughput screening of compounds against nonreplicating *Mycobacterium tuberculosis*. *Antimicrob. Agents Chemother.* **51**, 1380–1385 (2007).
39. Mak, P. A. *et al.* A high-throughput screen to identify inhibitors of ATP homeostasis in non-replicating *Mycobacterium tuberculosis*. *ACS Chem. Biol.* **7**, 1190–1197 (2012).
40. Wayne, L. G. *In vitro* model of hypoxically induced nonreplicating persistence of *Mycobacterium tuberculosis*. *Methods Mol. Med.* **54**, 247–269 (2001).
41. Lakshminarayana, S. B. *et al.* Comprehensive physicochemical, pharmacokinetic and activity profiling of anti-TB agents. *J. Antimicrob. Chemother.* **70**, 857–867 (2015).
42. Dhar, N. & McKinney, J. D. Microbial phenotypic heterogeneity and antibiotic tolerance. *Curr. Opin. Microbiol.* **10**, 30–38 (2007).
43. Mitchison, D. A. The search for new sterilizing anti-tuberculosis drugs. *Front. Biosci.* **9**, 1059–1072 (2004).
44. Franzblau, S. G. *et al.* Comprehensive analysis of methods used for the evaluation of compounds against *Mycobacterium tuberculosis*. *Tuberculosis (Edinb.)* **92**, 453–488 (2012).
45. Silva-Miranda, M. *et al.* High-content screening technology combined with a human granuloma model as a new approach to evaluate the activities of drugs against *Mycobacterium tuberculosis*. *Antimicrob. Agents Chemother.* **59**, 693–697 (2015).
46. Pethe, K. *et al.* Discovery of Q203, a potent clinical candidate for the treatment of tuberculosis. *Nat. Med.* **19**, 1157–1160 (2013).
47. Russell, D. G., Cardona, P. J., Kim, M. J., Allain, S. & Altare, F. Foamy macrophages and the progression of the human tuberculosis granuloma. *Nat. Immunol.* **10**, 943–948 (2009).
48. Harper, J. *et al.* Mouse model of necrotic tuberculosis granulomas develops hypoxic lesions. *J. Infect. Dis.* **205**, 595–602 (2012).
49. Datta, M. *et al.* Anti-vascular endothelial growth factor treatment normalizes tuberculosis granuloma vasculature and improves small molecule delivery. *Proc. Natl Acad. Sci. USA* **112**, 1827–1832 (2015).
50. Hawn, T. R., Shah, J. A. & Kalman, D. New tricks for old dogs: countering antibiotic resistance in tuberculosis with host-directed therapeutics. *Immunol. Rev.* **264**, 344–362 (2015).
51. Moore, E. M. & Lockwood, D. N. Treatment of visceral leishmaniasis. *J. Glob. Infect. Dis.* **2**, 151–158 (2010).
52. Sundar, S. & Chakravarty, J. An update on pharmacotherapy for leishmaniasis. *Expert Opin. Pharmacother.* **16**, 237–252 (2014).
53. Molina, I. *et al.* Randomized trial of posaconazole and benznidazole for chronic Chagas' disease. *N. Engl. J. Med.* **370**, 1899–1908 (2014).
54. De Rycker, M. *et al.* Comparison of a high-throughput high-content intracellular *Leishmania donovani* assay with an axenic amastigote assay. *Antimicrob. Agents Chemother.* **57**, 2913–2922 (2013).
55. Keenan, M. *et al.* Selection and optimization of hits from a high-throughput phenotypic screen against *Trypanosoma cruzi*. *Future Med. Chem.* **5**, 1733–1752 (2013).
56. Pena, I. *et al.* New compound sets identified from high throughput phenotypic screening against three kinetoplastid parasites: an open resource. *Sci. Rep.* **5**, 8771 (2015).
57. Don, R. & Ioset, J. R. Screening strategies to identify new chemical diversity for drug development to treat kinetoplastid infections. *Parasitology* **141**, 140–146 (2014).
58. Lipinski, C. A., Lombardo, F., Dominy, B. W. & Feeney, P. J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Deliv. Rev.* **46**, 3–26 (2001).
59. Gilbert, I. H. Drug discovery for neglected diseases: molecular target-based and phenotypic approaches. *J. Med. Chem.* **56**, 7719–7726 (2013).
60. Ansari, M. T. *et al.* Malaria and artemisinin derivatives: an updated review. *Mini Rev. Med. Chem.* **13**, 1879–1902 (2013).
61. Clark, R. L. Embryotoxicity of the artemisinin antimalarials and potential consequences for use in women in the first trimester. *Reprod. Toxicol.* **28**, 285–296 (2009).
62. Toovey, S. Safety of artemisinin antimalarials. *Clin. Infect. Dis.* **42**, 1214–1215 (2006).
63. Braselmann, S. *et al.* R406, an orally available spleen tyrosine kinase inhibitor blocks fc receptor signaling and reduces immune complex-mediated inflammation. *J. Pharmacol. Exp. Ther.* **319**, 998–1008 (2006).
64. Oberlies, N. H. & Kroll, D. J. Camptothecin and taxol: historic achievements in natural products research. *J. Nat. Prod.* **67**, 129–135 (2004).
65. Tsang, C. K., Qi, H., Liu, L. F. & Zheng, X. F. Targeting mammalian target of rapamycin (mTOR) for health and diseases. *Drug Discov. Today* **12**, 112–124 (2007).
66. Borel, J. F. History of the discovery of cyclosporin and of its early pharmacological development. *Wien Klin. Wochenschr.* **114**, 433–437 (2002).
67. Slingsby, B. T. & Kurokawa, K. The Global Health Innovative Technology (GHIT) fund: financing medical innovations for neglected populations. *Lancet Glob. Health* **1**, e184–185 (2013).
68. Holmes, D. The GHIT fund shows its cards. *Nat. Rev. Drug Discov.* **12**, 894 (2013).
69. Crump, A. & Omura, S. Ivermectin, 'wonder drug' from Japan: the human use perspective. *Proc. Jpn Acad., Ser. B* **87**, 13–28 (2011).
70. Kita, K., Shiomi, K. & Omura, S. Advances in drug discovery and biochemical studies. *Trends Parasitol.* **23**, 223–229 (2007).
71. Omura, S. & Crump, A. The life and times of ivermectin - a success story. *Nat. Rev. Microbiol.* **2**, 984–989 (2004).
72. Strader, C. R., Pearce, C. J. & Oberlies, N. H. Fingolimod (FTY720): a recently approved multiple sclerosis drug based on a fungal secondary metabolite. *J. Nat. Prod.* **74**, 900–907 (2011).
73. Endo, A. A historical perspective on the discovery of statins. *Proc. Jpn Acad. Ser. B Phys. Biol. Sci.* **86**, 484–493 (2010).
74. Tobert, J. A. Lovastatin and beyond: the history of the HMG-CoA reductase inhibitors. *Nat. Rev. Drug Discov.* **2**, 517–526 (2003).
75. Baell, J. B. & Holloway, G. A. New substructure filters for removal of pan assay interference compounds (PAINS) from screening libraries and for their exclusion in bioassays. *J. Med. Chem.* **53**, 2719–2740 (2010).
76. McGovern, S. L., Caselli, E., Grigorieff, N. & Shoichet, B. K. A common mechanism underlying promiscuous inhibitors from virtual and high-throughput screening. *J. Med. Chem.* **45**, 1712–1722 (2002).
77. Baell, J. & Walters, M. A. Chemistry: chemical con artists foil drug discovery. *Nature* **513**, 481–483 (2014).
78. Burrows, J. N., Leroy, D., Lotharius, J. & Waterson, D. Challenges in antimalarial drug discovery. *Future Med. Chem.* **3**, 1401–1412 (2011).
79. Ruecker, A. *et al.* A male and female gametocyte functional viability assay to identify biologically relevant malaria transmission-blocking drugs. *Antimicrob. Agents Chemother.* **58**, 7292–7302 (2014).
80. Orme, I. Cellular and genetic mechanisms underlying susceptibility of animal models to tuberculosis infection. *Novartis Found. Symp.* **217**, 112–117; discussion 117–119 (1998).
81. Priest, B. T., Bell, I. M. & Garcia, M. L. Role of hERG potassium channel assays in drug development. *Channels (Austin)* **2**, 87–93 (2008).
82. Riss, T. L. *et al.* Cell viability assays. *Assay Guidance Manual* [online], <http://www.ncbi.nlm.nih.gov/books/NBK53196/> (2004).

## Competing interests statement

The authors declare no competing interests.

## FURTHER INFORMATION

Drugs for Neglected Diseases initiative: <http://www.dndi.org>

Global Health Innovative Technology (GHIT) Fund:

<http://www.ghitfund.org>

Japanese Pharmaceutical Manufacturers Association:

<http://www.jpma.or.jp>Medicines for Malaria Venture: <http://www.mmv.org/>TB Alliance: <http://www.tballiance.org/>WHO website: <http://www.who.int/tb>

ALL LINKS ARE ACTIVE IN THE ONLINE PDF

## A community-based approach to new antibiotic discovery

Matthew A. Cooper

The Community for Open Antimicrobial Drug Discovery aims to tap into the potential of the millions of compounds distributed around laboratories globally to be a source of new antibiotic leads by offering free screening for antimicrobial properties, with no strings attached.

There has been much written on the ‘superbug crisis’ over the past decade. Commonly cited causes include the over-use of antibiotics in humans and animals and the exit of many companies from antibacterial drug research and development (R&D) owing to the challenging market economics<sup>1</sup>. However, there is another important cause — it is simply getting more and more difficult to find new antibacterial drugs. Most of the current arsenal of antibacterial drugs originated from screening of natural products in the 1950s–1970s. However, the ‘low-hanging fruit’ from this strategy appears to have been exhausted, and one alternative — target-based screening using knowledge of bacterial genomics — has failed to replenish the drug pipeline<sup>2</sup>. Consequently, there has been a revival in interest in establishing antimicrobial screening platforms to interrogate previously untapped sources of chemical diversity.

### New sources of chemical diversity

One way to gain access to new chemical diversity that could be highly relevant for antibacterial drug discovery is to isolate novel natural products produced by microbes; for example, from bacteria that have previously been ‘unculturable’<sup>3</sup>. However, when a new natural product with antibacterial activity is found, it is not easy to isolate and scale up the production of sufficient material needed for preclinical development. Furthermore, antibiotic-resistance mechanisms to a novel natural product will presumably already be present in the bacterial community, given the need for the organism producing the antibiotic to protect itself against the effects of the antibiotic and the defensive mechanisms evolved by other bacteria to the agent.

The good news is that we have already made many other molecules that have not been encountered by microbes before, and that have yet to be tested for antimicrobial activity. At the time of writing, 94 million compounds had been deposited in the Chemical Abstract Service (CAS) Registry; 80 million of these were organic compounds with no metal ion and a molecular mass

of <1,500 Da. Chemists around the world synthesize thousands of new organic compounds each year, but most of these compounds never leave their laboratories. Individually, such compounds may have little value, but taken collectively as part of a ‘global collection’ of unusual synthetic molecules, they could be much more useful. Chemists — and academic chemists in particular — make molecules for many reasons. Most chemists do not work in drug design, but are instead interested in the development of new synthetic methods and novel structures, and so are unconstrained by concerns about ‘drug-like’ properties. By contrast, pharmaceutical companies hold large collections of compounds that are curated to maximize the likelihood of oral bioavailability, and biased towards major human target families, such as G protein-coupled receptors and kinases. Furthermore, antibiotics occupy very diverse chemical space, and many are ‘suicide inhibitors’ of microbial enzymes and often possess reactive groups that are normally filtered out<sup>4</sup> in corporate collections.

So, how many of the 80 million organic compounds in the CAS Registry might have the potential to be starting points for new antibacterial drugs? Applying a simple filter for antibacterial-like properties<sup>5</sup> (calculated logP between –10 and 2, and molecular mass <1,200 Da) to this group yielded 29 million compounds. The majority of these compounds are not commercially available, and most will never have been screened for antibacterial activity. Importantly, however, any compounds within this subset would have a known synthetic route, facilitating the scaling up of production of validated hits for further study and the exploration of analogues with improved properties.

Given these initial considerations, the key question is: are there novel, high-quality antimicrobial hits among the millions of synthetic organic compounds distributed around laboratories globally? In 2015, we launched the [Community for Open Antimicrobial Drug Discovery \(CO-ADD\)](#), which is funded by the Wellcome Trust and The University of Queensland, to engage the worldwide chemistry community in an effort to answer this question.

Matthew A. Cooper is the Director of the Community for Open Antimicrobial Drug Discovery, Institute for Molecular Bioscience, The University of Queensland, Brisbane, Queensland 4072, Australia.  
e-mail: [m.cooper@uq.edu.au](mailto:m.cooper@uq.edu.au)  
doi:10.1038/nrd4706  
Published online  
7 August 2015



### A community approach: CO-ADD

CO-ADD is an open-access facility, in that any chemist in the world can have their compounds tested for antimicrobial potency and cytotoxicity, free of charge and with no encumbrance on intellectual property (IP).

Compounds submitted to CO-ADD undergo a primary screen in a 384-well format against representatives of 5 key bacterial pathogens (*Escherichia coli*, *Klebsiella pneumoniae*, *Acinetobacter baumannii*, *Pseudomonas aeruginosa* and methicillin-resistant *Staphylococcus aureus* (MRSA), as well as the fungi *Cryptococcus neoformans* and *Candida albicans*. Membrane-deficient and efflux pump-impaired *E. coli* mutants are also assayed to enhance our knowledge surrounding antibiotic activity, penetration and inactivation. Subsequent hit confirmation includes dose–response antimicrobial assays, as well as counter-screening for adverse effects (using cytotoxicity, critical micelle concentration and membrane disruption assays), which is then followed by a cascade of assays for hit validation. Here, the hit compounds are assayed against a broader panel of multidrug-resistant clinical isolates and reference strains and tested in the presence of serum and lung surfactant, and then assessed for blood haemolysis, microsomal and plasma stability and drug–protein binding.

Together, these data support nomination of a validated hit and form the basis for optimization towards a candidate antibiotic. At this stage, it is not possible to conduct lead optimization for free. However, such activities could potentially be funded through schemes such as the Wellcome Trust Seeding Drug Discovery awards, the US National Institute of Allergy and Infectious Diseases R21/R33 awards and the European Innovative Medicines Initiative ENABLE programme.

One key aspect of the CO-ADD programme is the use of standardized testing based on well-defined, consistent protocols. Many other fields of drug discovery use standard assays and instrumentation; however, in antimicrobial discovery, a multitude of assays (including disc diffusion, Etest and broth or agar dilution assays), strains, media and plate types are employed, meaning a direct comparison of results between laboratories is challenging. Clinical and Laboratory Standards Institute (CLSI)-compliant, standardized conditions and reference strains will be used throughout the programme, so that researchers will finally be able to compare thousands of compounds assayed under identical conditions. As more compounds are screened, the collated data, which will be made publicly available for both active and non-active compounds, will become increasingly valuable for gaining understanding not just of novel chemotypes associated with antibacterial activity, but also of characteristics that influence other key properties, such as penetration into bacterial cells and avoidance of efflux.

CO-ADD is a hybrid of existing successful models: those that provide access to facilities (for example, GlaxoSmithKline's Open Lab Foundation at Tres Cantos, and Eli Lilly's Open Lab), and discovery consortia (such as the US National Institutes of Health Molecular Libraries Program, the Tropical Disease Initiative, Collaborative Drug Discovery and EU-OPENSREEN).

CO-ADD is differentiated from these by its specific focus on antimicrobial screening and discovery, with an extremely low barrier to participant entry. There are no limitations on geography, nor expert panels that review molecules for 'suitability' to screen. The approach is fully inclusive — if the compound is soluble in water or dimethyl sulfoxide, not radioactive, not pyrophoric and not an illicit drug, CO-ADD will screen it.

### Challenges and outlook

In the golden age of antibiotic discovery, most scientists would collaborate openly with each other. Now university technology transfer offices are reluctant to allow researchers to transfer compounds, assays or reagents without non-disclosure agreements and material transfer agreements (MTAs), which can take many months to agree — even for limited, short experiments. CO-ADD provides simple terms and conditions online and a signature-free MTA download if needed, which together warrant that the provider of the compound retains all rights to the compound, assay results and IP. Furthermore, researchers are initially only asked to provide the molecular mass of the compound, and a structural fingerprint if possible. CO-ADD participants have 18 months to patent and develop their compound before they are asked to make structures and results available to the open-access database, and if 18 months is too short, this period can be extended on request. Hopefully, this compromise will lead to a valuable repository of structure–activity and structure–toxicity relationships, while allowing sufficient time for IP related to promising compounds to be obtained in order for commercially focused clinical development to progress.

There is also a practical barrier to entry — the simple act of weighing out and shipping a compound. Here, the CO-ADD kick-start programme provides financial incentives to labs to defray costs associated with sample collection and shipping. Our outreach strategy is raising awareness and building a network of advocates through conferences and presentations globally, and the programme has been received with enthusiasm by individuals, laboratories and university and institute collections.

Very few completely new classes of antibacterial or antifungal drugs have been approved in the past two decades. CO-ADD is just a small part of what we hope will be a renaissance in new R&D and business models and technologies that together can provide a sustainable solution to the superbug crisis.

1. Cooper, M. A. & Shlaes, D. A. Fix the antibiotics pipeline. *Nature*, **472**, 32 (2011).
2. Payne, D. J. *et al.* Drugs for bad bugs: confronting the challenges of antibacterial discovery. *Nat. Rev. Drug Discov.* **6**, 29–40 (2007).
3. Lewis, K. Platforms for antibiotic discovery. *Nat. Rev. Drug Discov.* **12**, 371–387 (2013).
4. Shushko, I. *et al.* ToxAlerts: a web server of structural alerts for toxic chemicals and compounds with potential adverse reactions. *J. Chem. Inf.* **52**, 2310–2316 (2012).
5. O'Shea, R. & Moser, H. E. Physicochemical properties of antibacterial compounds: implications for drug discovery. *J. Med. Chem.* **51**, 2871–2878 (2008).

#### Competing interests statement

The author declares no competing interests.

#### FURTHER INFORMATION

CO-ADD: <http://www.co-add.org/>

ALL LINKS ARE ACTIVE IN THE ONLINE PDF

## OPINION

# Pioneering government-sponsored drug repositioning collaborations: progress and learning

Donald E. Frail, Madeleine Brady, K. Jane Escott, Alison Holt, Hitesh J. Sanganeer, Menelas N. Pangalos, Chris Watkins and Craig D. Wegner

**Abstract** | A new model for translational research and drug repositioning has recently been established based on three-way partnerships between public funders, the pharmaceutical industry and academic investigators. Through two pioneering initiatives — one involving the Medical Research Council in the United Kingdom and one involving the National Center for Advancing Translational Sciences of the National Institutes of Health in the United States — new investigations of highly characterized investigational compounds have been funded and are leading to the exploration of known mechanisms in new disease areas. This model has been extended beyond these first two initiatives. Here, we discuss the progress to date and the unique requirements and challenges for this model.

Drug repositioning — also referred to as repurposing, reprofiling, rescue, or indications discovery — is the process of identifying a new use for an existing drug or drug candidate in an indication outside the scope of the original indication. The case was made a decade ago that drug repositioning is one strategy to address the declining productivity of the pharmaceutical industry, as starting with a well-characterized compound decreases the duration of clinical development and could reduce attrition owing to issues such as poor pharmacokinetics or insufficient safety<sup>1</sup>. Furthermore, well-characterized clinical-stage compounds can be used to investigate novel disease hypotheses in human studies, which is important given the limitations of studies in animal models<sup>2</sup>, including the lack of predictivity of the efficacy of compounds in subsequent clinical trials and issues with reproducibility (see the [Nature focus on challenges in reproducible research](#)). Making such compounds more widely available helps to drive hypothesis creation and can lead to broader testing in humans.

Multiple collaborations have since been established to enable the investigation of scientific advances within academia using drugs and drug candidates from industry. For example, in 2010, Pfizer's Indications Discovery Unit and the Washington University School of Medicine formed a partnership that provided access to

proprietary data for a large portfolio of active and discontinued Pfizer drug candidates and allowed investigators to propose and collaborate on preclinical or clinical studies to investigate new uses<sup>3,4</sup>. Around the same time, the UK Medical Research Council (MRC) was seeking to enable a greater understanding of the mechanisms of human disease through experimental medicine studies enabled by access to high-quality compounds from industry. Although this strategic intent is distinct from that of a drug repositioning effort, the two complement each other. Meanwhile, Francis Collins, Director of the US National Institutes of Health (NIH), advocated for a more comprehensive repositioning programme that harnessed the strength of various stakeholders<sup>5</sup>. The NIH convened a roundtable of leading representatives from academia, the US government and private sector research and development (R&D) to explore such strategies, including one in which pharmaceutical companies would create a pool of compounds for further investigation by academia through a government grant programme<sup>6</sup>. Thus, in 2011 the MRC, in partnership with AstraZeneca, implemented the Mechanisms for Human Diseases Initiative, and in 2012 the NIH, through the National Center for Advancing Translational Sciences (NCATS), implemented the [Discovering New Therapeutic Uses for Existing Molecules](#) initiative.

The announcements of these programmes created some controversy, with some questioning the role of the NIH and NCATS as drug developers and the value of drug repositioning (for example, see REF. 6 and further discussion below). Given the timeframes of drug development, it may be several more years before the full value of these programmes is clear. Here, we make an interim assessment of the progress to date, discussing the unique requirements and challenges encountered with these programmes, as well as early indicators of success.

## Programme characteristics

**The MRC Mechanisms of Human Disease Initiative.** This programme, a partnership between the MRC and AstraZeneca that was launched in 2011, provided academic researchers with unprecedented access to a high-quality collection of clinical and preclinical AstraZeneca compounds in order that they could propose new research into human disease mechanisms and the development of potential therapeutic interventions. The two partners had different but complementary motives. For the MRC, the initiative supported the [MRC Translational Research Strategy](#), which has a strong emphasis on experimental medicine research to understand the biology of human disease and includes, where appropriate, preclinical studies using model systems. Successful preclinical studies could stimulate further pursuit of mechanistically related compounds in the clinic. The development of potential therapeutic interventions was not a primary goal, but it was hoped that successful studies and an increased understanding of the mechanisms involved in human disease, which would go into the public domain via peer-reviewed publications, would lead to the development of new medicines for patients by AstraZeneca or other companies. For AstraZeneca, the programme also provided the opportunity to build stronger relationships with members of the UK academic community with knowledge across a broad range of diseases.

Criteria were established for identifying which AstraZeneca compounds could be made available for external research. The focus was on identifying either development compounds that were discontinued and considered suitable for additional clinical studies (including confidence that target engagement is achievable), which were offered for clinical and preclinical proposals, or development compounds that were no

longer deemed suitable for clinical studies based on available clinical data (for example, lack of target coverage or sufficient safety) or limitations identified preclinically (for example, emerging preclinical long-term safety data), which were offered for preclinical proposals only.

AstraZeneca filtered through over 450 compounds that had been nominated for clinical development and identified a total of 22 suitable compounds (see [Supplementary information S1 \(table\)](#)). These compounds had been extensively characterized, including their potency, selectivity, pharmacology, complete pharmacokinetic and safety packages, target engagement and previous use in humans. For each compound, internal data were reviewed and summaries of the most relevant information were developed to enable investigators to craft a suitable proposal (see [Supplementary information S2 \(table\)](#)). This information was posted on the MRC website (that is, in the public domain) to attract the most innovative ideas from MRC-eligible investigators. At the time, this compendium of information was the largest single public source of what was proprietary information regarding efficacy, safety and other information on discontinued development compounds.

The MRC initiated a call for 'concept proposals', directing investigators to the compounds on the website. Over the 8-week open call period, more than 100 proposals were submitted from 37 different UK institutions, across all compounds and spanning a range of disease areas. The geographic diversity and breadth of ideas indicated that the concept of crowdsourcing was successful.

About half of the proposals fell within disease areas of focus for AstraZeneca, with the remainder falling into areas outside core therapeutic areas of interest, although whether the proposal fell within an area of interest to AstraZeneca was not a criterion used for review. A committee comprised of senior UK scientists convened by the MRC provided an initial review of the proposals for scientific merit, and independently, AstraZeneca provided a review of both the feasibility and suitability of the compound for the proposed investigations and the novelty of the studies. AstraZeneca was able to review these proposals owing to a confidentiality disclosure agreement (CDA) between AstraZeneca and the MRC, which covered these activities.

Informed by these reviews, a joint MRC–AstraZeneca committee then identified 25 proposals to advance to full proposal. The full proposals were collaboratively

developed by the UK investigators and AstraZeneca scientists, and a champion of the proposal from AstraZeneca was required to proceed. Given the innovative funding approach taken, the collaboration between UK investigators and AstraZeneca scientists was essential for the development and pursuit of highly competitive and scientifically robust proposals. The full proposal development process was intense, with some complex clinical investigations being considered within short timelines. In all cases, AstraZeneca researchers were named co-applicants on the proposals, which were led by an academic. The full applications were considered by a specially convened MRC committee who were informed by international peer review comments; there was no AstraZeneca representation in the committee to ensure funding decisions were based solely on scientific merit and feasibility, without any commercial bias.

In total, 15 collaborative proposals — 7 preclinical and 8 clinical — were funded by the MRC and are now underway ([TABLE 1](#)). The collection of studies is highly diverse and each study is for a different indication, ranging from common disorders (for example, Alzheimer disease) to orphan diseases (for example, muscular dystrophy) and including indications outside of AstraZeneca's core areas of focus. Funding for the studies is provided to the investigators by the MRC (no MRC funds are provided to AstraZeneca). AstraZeneca, in addition to providing collaborative insight and suggestions to the full proposals, is responsible for providing the necessary drug supply and documents to support the regulatory and ethics committee filings by the investigators, as well as coordination of any adverse events when a compound is the subject of more than one study. The clinical studies are sponsored by the investigator, who is therefore ultimately responsible for the conduct of the study.

**NIH–NCATS Discovering New Therapeutic Uses for Existing Molecules.** The pilot programme from NCATS was initiated in 2012 (see [Further information](#)). This programme matched NIH-funded researchers with a selection of 57 compounds previously discontinued from development (see [Supplementary information S1 \(table\)](#)). In this case, multiple companies were involved, with Abbott (supplying 3 compounds), Bristol-Myers Squibb (3), GlaxoSmithKline (4), Johnson & Johnson (3) and Sanofi (10) joining the founding companies of AstraZeneca (14), Eli Lilly (4) and

Pfizer (17). Although similar in essence to the MRC programme, there were also substantial differences ([TABLE 2](#)):

- Purpose of the programmes. The primary goal of the MRC programme was to investigate mechanisms of human disease. Therefore, the MRC programme included preclinical studies, early concept-testing human studies and/or Phase II proof-of-concept studies. By contrast, the NCATS programme required the full proposal to contain a statistically powered Phase II proof-of-concept study. Neither the MRC programme nor the NIH programme was a global programme — each targeted only eligible investigators in the United Kingdom or the United States, respectively, although both were essential in building the foundation for a global programme as is discussed below (see also the [AstraZeneca Open Innovation](#) website).
- Compound criteria. The NCATS programme required compounds to have prior evidence of target coverage and manageable tolerability in humans, to potentially enable confident hypothesis testing in a new indication. The MRC programme also included compounds suitable for only preclinical use with evidence of potency, selectivity and exposure (typically by the oral route) in preclinical models. In both programmes, compounds were no longer in active development (that is, they were discontinued) to avoid any concern that public funding might be supplementing an ongoing commercial development objective.
- Review process. In the NCATS programme, the review of the concept proposals did not involve the companies, and proposals that were not selected for full proposal were not seen by the industry partners. The pharmaceutical companies were engaged under a CDA for those proposals selected for full development and could deny support of the proposal, thereby preventing full proposal submission. As with the MRC programme, the final funding decisions were made without company input.
- Template agreements. NCATS required each company to prepare and publicly post online template CDAs and collaborative research agreements (CRAs) at the time of the announcement of the programme, enabling early review by technology transfer offices and rapid implementation of such agreements. In addition, NCATS required a signed



Table 1 | **Government-sponsored collaborative drug repositioning projects**

Project focus*	Collaborators, institution and industry partner	New indication (original indication)	Type of project
<b>MRC-funded projects</b>			
Saracatinib (AZD0530) as a novel analgesic for cancer-induced bone pain	• Dr D. Andrew, University of Sheffield, UK • AstraZeneca	Cancer-induced bone pain (solid tumour)	Preclinical and clinical
Assessing the therapeutic efficacy of an 11 $\beta$ HSD1 inhibitor (AZD4017) in idiopathic intracranial hypertension	• Dr A. Sinclair, University of Birmingham, UK • AstraZeneca	Idiopathic intracranial hypertension (diabetes and obesity)	Clinical
Evaluation of the selective endothelin A-receptor antagonist zibotentan (AZD4054) as a treatment for renal disease in systemic sclerosis (scleroderma)	• Professor C. Denton, University College London, UK • AstraZeneca	Renal scleroderma (prostate cancer)	Clinical
Phase II study of the impact of a selective 11 $\beta$ HSD1 inhibitor (AZD4017) on biochemical markers of bone turnover in post-menopausal osteopaenia	• Professor P. Stewart, University of Leeds, UK • AstraZeneca	Post-menopausal osteopaenia (diabetes and obesity)	Clinical
The role of GABA <sub>B</sub> receptor mechanisms in chronic cough (using AZD3355)	• Professor J. Smith, University of Manchester, UK • AstraZeneca	Chronic cough (gastroesophageal reflux disease)	Clinical
Exploring GABA <sub>A</sub> $\alpha$ 2,3 signalling as novel therapy for peripheral neuropathies and primary dystonias (using AZD7325)	• Professor M. Koltzenburg, University College London, UK • AstraZeneca	Dystonia or neuropathy (anxiety)	Clinical
SRC inhibitors (AZD0530) as potential antipsychotics: human testing with psilocybin	• Professor D. Nutt, Imperial College London, UK • AstraZeneca	Psychosis (solid tumour)	Clinical
A new paradigm for testing pathway tractability in lung disease (using an MMP9 and MMP12 inhibitor (AZD1236))	• Dr N. Hirani, University of Edinburgh, UK • AstraZeneca	Idiopathic pulmonary fibrosis (chronic obstructive pulmonary disease)	Preclinical and clinical
The role of MMP inhibitors (AZD1236) in ameliorating muscular dystrophy	• Professor D. Wells, Royal Veterinary College, London, UK • AstraZeneca	Muscular dystrophy (chronic obstructive pulmonary disease)	Preclinical
Investigating ATP regulation and P2X7 blockade (AZ11657312) in acute renal injury and its long-term complications	• Professor R. Unwin, University College London, UK • AstraZeneca	Acute kidney injury (rheumatoid arthritis)	Preclinical
Efficacy of saracatinib (AZD0530) in treatment of chronic otitis media in preclinical mouse models	• Dr M. Cheeseman, University of Edinburgh, UK • AstraZeneca	Chronic otitis media (solid tumour)	Preclinical
Evaluation of AZD1080 (GSK3 $\beta$ inhibitor) in a preclinical mouse model of motor neuron disease	• Dr R. Mead, University of Sheffield, UK • AstraZeneca	Amyotrophic lateral sclerosis (Alzheimer disease)	Preclinical
Endothelin-1-mediated reduction of cerebral blood flow in Alzheimer disease: therapeutic potential of zibotentan (AZD4054)	• Professor S. Love, University of Bristol, UK • AstraZeneca	Alzheimer disease (prostate cancer)	Preclinical
GSK3 as a multifunctional target for glioblastoma treatment; hitting multiple tumour hallmarks with a single drug (AZD2858)	• Dr S. Short, University of Leeds, UK • AstraZeneca	Glioblastoma (Alzheimer disease)	Preclinical
<b>NIH-NCATS-funded projects</b>			
The efficacy and safety of a selective oestrogen receptor- $\beta$ agonist (LY500307)	• Dr A. Breier, Indiana University, Indianapolis, USA • Eli Lilly & Co.	Schizophrenia (benign prostatic hyperplasia)	Clinical
FYN inhibition by AZD0530 for Alzheimer disease	• Professor S. Strittmatter, Dr H. Nygaard and Professor C. Van Dyck, Yale University, New Haven, Connecticut, USA • AstraZeneca	Alzheimer disease (solid tumour)	Clinical
Medication development of a novel therapeutic for smoking cessation	• Dr H. Brunzell, Virginia Commonwealth University, Richmond, Virginia, USA • Dr K. Perkins, University of Pittsburgh, Pennsylvania, USA • Janssen Research & Development, LLC	Smoking cessation (psoriasis and rheumatoid arthritis)	Clinical

Table 1 (cont.) | **Government-sponsored collaborative drug repositioning projects**

Project focus*	Collaborators, institution and industry partner	New indication (original indication)	Type of project
A novel compound for alcoholism treatment: a translational strategy	<ul style="list-style-type: none"> <li>• Professor F. Akhlaghi, University of Rhode Island, Kingston, New York, USA</li> <li>• Dr L. Leggio, National Institute on Alcohol Abuse and Alcoholism and National Institute on Drug Abuse, Bethesda, Maryland, USA</li> <li>• Pfizer</li> </ul>	Alcoholism (type 2 diabetes)	Clinical
Partnering to treat an orphan disease: Duchenne muscular dystrophy	<ul style="list-style-type: none"> <li>• Dr K. Wagner, Kennedy Krieger Institute, Baltimore, Maryland, USA</li> <li>• Dr S. Froehner, University of Washington, Seattle, USA</li> <li>• Sanofi</li> </ul>	Duchenne muscular dystrophy (not reported)	Clinical
Reuse of ZD4054 for patients with symptomatic peripheral artery disease	<ul style="list-style-type: none"> <li>• Dr B. Annex, University of Virginia, Charlottesville, USA</li> <li>• AstraZeneca</li> </ul>	Peripheral arterial disease (prostate cancer)	Clinical
Therapeutic strategy for lymphangioleiomyomatosis (AZD0530 (saracatinib))	<ul style="list-style-type: none"> <li>• Dr T. Eissa, Baylor College of Medicine, Houston, Texas, USA</li> <li>• AstraZeneca</li> </ul>	LAM and TSC (solid tumour)	Clinical
Therapeutic strategy to slow progression of calcific aortic valve stenosis	<ul style="list-style-type: none"> <li>• Dr J. Miller, Dr M. Enriquez-Sarano and Dr H. Schaff, Mayo Clinic, Rochester, New York, USA</li> <li>• Sanofi</li> </ul>	Calcific aortic valve stenosis (not reported)	Clinical
Translational neuroscience optimization of GlyT1 inhibitor	<ul style="list-style-type: none"> <li>• Dr J. Krystal, Yale University, New Haven, Connecticut, USA</li> <li>• Pfizer</li> </ul>	Cognitive deficits in schizophrenia (schizophrenia)	Clinical

\*Compound codenames are provided where possible. 11 $\beta$ HSD1, 11 $\beta$ -hydroxysteroid dehydrogenase type 1; GABA,  $\gamma$ -aminobutyric acid; GlyT1, glycine transporter 1; GSK3, glycogen synthase kinase 3; LAM, lymphangioleiomyomatosis; MMP, matrix metalloproteinase; MRC, UK Medical Research Council; NCATS, National Center for Advancing Translational Sciences; NIH, National Institutes of Health; TSC, tuberous sclerosis complex.

CRA to be submitted with the final full proposal, which provided a deadline for any negotiations, whereas CRAs for the MRC programme were negotiated after the awards were granted to minimize the administrative burden of negotiating agreements with companies that might not have been awarded funding. The MRC process was somewhat facilitated by existing templates — the [NIHR–MRC model Industry Collaborative Research Agreement](#) (mICRA) and the [UK Government Lambert agreements](#) — which were further adapted for the purposes of this specific initiative. Despite the existence of the template agreements, some negotiation was still required.

Ultimately, nine clinical proposals were funded by the NIH, as summarized in TABLE 1 (see REF. 7 for an overview of this programme).

**National Research Program for Biopharmaceuticals.** In 2013, a similar programme was established between AstraZeneca and the National Research Program for Biopharmaceuticals (NRPB) in Taiwan to facilitate translational research locally. This programme combines elements

of the NIH–NCATS programme (for example, posting of a template CRA) and the MRC programme (both clinical and preclinical-only proposals supported). However, for the first time in these relationships, compounds actively being pursued in development ('live' compounds) at AstraZeneca were included, thereby setting a new precedent in this type of setting (see Supplementary information S1 (table)). One clinical and two preclinical projects were funded and are in progress. Additional proposals resulting from the networking and relationships established under this programme are now under collaborative discussions between the investigator, AstraZeneca and the NRPB.

#### Responses to the programmes

The announcement of the MRC programme was met with enthusiasm and seen as an exciting, unprecedented opportunity by UK investigators. When the NIH–NCATS programme was announced, it too was seen as an unprecedented opportunity, although there was some scepticism and criticism.

Some criticized the allocation of public funds to investigate company compounds and questioned the return to the public sector. However, academic clinical

investigations, including those involving company compounds, are routinely supported by public funds, and these programmes provided access to compounds for mechanisms that have not previously been available. John LaMattina, former President of Research and Development at Pfizer, was one of many who criticized the involvement of academic investigators in drug development<sup>6</sup>. Such critics argued that the NIH (broadly used as a term to apply to academic researchers) lacks the experience required to develop drugs. Importantly, however, the intent of the programmes was to identify new uses of existing compounds and known mechanisms, not to develop drugs through to regulatory approval. Furthermore, previous experience of efforts to test company compounds in different tumour types, supported by the US National Cancer Institute (NCI), demonstrated that academic investigators are capable of achieving this goal. Examples in which NCI involvement in the early stages of development ultimately resulted in new medicines include cisplatin for the treatment of testicular, ovarian and lung cancer, and paclitaxel and fludarabine phosphate for the treatment of several cancers and lymphoma, respectively. In addition, projects supported by the

NCATS programme required company collaboration to develop the final proposal, thus combining the experience and knowledge of both academic and industry investigators.

Another criticism was that company scientists have already exhaustively considered alternative indications for their compounds. This is not always true, as companies typically have therapeutic areas of focus and only invest time and money in those. The indications considered by these programmes were unlimited and utilized the broad expertise lying outside pharmaceutical companies. More importantly, science continuously progresses with new discoveries, and some of these discoveries could lead to a new use for an existing compound, for example, unpublished data from investigators linking a target to a disease. Furthermore, the totality of the available data may not be sufficient for investment by the company, and additional preclinical or clinical data may change interest levels.

There was also criticism of these programmes because the compound structures were not initially disclosed<sup>8</sup>. This criticism primarily came from academic groups doing *in silico* drug repositioning, in which structural features of one compound are found to be similar to those of a drug with a known effect that acts through a different target (off-target activity). However, the purpose of the programmes was hypothesis-based repositioning with defined mechanisms of action, not to generate new hypotheses for an individual molecule working through off-target biology or other repositioning approaches. It was concluded that non-hypothesis-based investigations were not within the scope of these programmes and that the extensive compound information provided to investigators was sufficient to meet the goals of the programme, thus compound structures were not required. Nevertheless, Southan *et al.* did take the initiative to search publicly available databases to identify most, if not all, of the compound structures in the NIH–NCATS programme, although the accuracy was not confirmed by the companies (see REF. 8 and the [Southan figshare page](#)).

Caution and scepticism was also evident within pharmaceutical companies. Many clinical compounds, although listed as discontinued after failing to improve on the current standard of care in their initial Phase II indication, remain under preclinical evaluation for alternative indications. This status typically persists for a few years, during which time R&D leaders can be reluctant to allow the compound to be tested externally. By the time this internal

deliberation is exhausted, interest in the compound and its mechanism of action, both within and outside the company, and the patent life have diminished. In addition, as R&D budgets have decreased over the past decade many companies have narrowed their therapeutic area focus, leading to hesitation by some to support investigator-sponsored trials in non-core therapy areas. Budget constraints can therefore make it challenging for companies to supply the clinically formulated drug (and matched placebo) substance, updated regulatory documents (for example, the Investigator's Brochure and Chemical, Manufacturing and Controls section) and scientific and clinical compound-specific advice for disease areas that are not a high priority. Recent activities, described below, indicate that more and more companies have overcome these initial concerns.

### Unique requirements and challenges

Clinical trials generally fall into one of two groups: company-sponsored studies or investigator-sponsored studies. To date, investigator-sponsored studies, regardless of the funder, almost always use live compounds that are in development or on the market. In these circumstances, existing project teams provide the required compound insight, regulatory document updates, safety database access and so on, and they are ultimately the decision makers for whether a given study should be run. For a discontinued compound, given that the project team has often been disbanded, there are unique challenges that require new ways of working to be established.

**Compound selection.** The initial MRC and NIH programmes were constructed to include discontinued compounds as these provided the easiest initial path to develop and implement such a groundbreaking activity. However, the status of a compound can be dynamic. As an example, one discontinued compound, a hormone modulator, was repositioned within the company for polycystic ovarian syndrome and a Phase II study was initiated while the MRC programme was being established. In this case, one MRC proposal was similar to the internal programme and led to a separate collaboration with a leading investigator in the area that focused on key scientific questions for the programme. It is therefore important in such collaborations to be flexible and enable a compound to be used in more than one collaboration or to be re-evaluated for development within the company.

The following general criteria were established by AstraZeneca and agreed to by the MRC and NCATS to select compounds for these publicly sponsored collaborative drug-repositioning programmes:

- Clinical and/or preclinical evidence of potency, selectivity and exposure supporting target coverage was required to ensure conclusive testing of a novel mechanism-driven hypothesis.
- For clinical proof-of-concept testing, sufficient patient safety to support further development was required. This involved substantial analysis and judgment in the context of the indication, length of study or target patient population. The length of supporting safety studies was highlighted to alert investigators of the acceptable length of proposed clinical studies. Investigators could include longer-term toxicology studies in their proposals if necessary.
- Reasonable cross-species activity and suitability for dosing in animals was required for all compounds made available for non-clinical studies.
- Patent life was not required. However, sufficient remaining patent life, new intellectual property or regulatory data exclusivity would probably be required to advance positive findings further.
- There had to be no other relevant commitments or complex legal agreements (for example, a compound is not partnered with or licensed to a third party).
- For compounds to be used for clinical studies, enough drug substance to support a reasonably sized clinical trial had to be available. The same route of administration as previously used was required to avoid time, cost and attrition risk associated with new formulation development, safety studies and human pharmacokinetic studies. Doses were typically limited to those supported by the existing data, though in rare cases additional preclinical data were obtained as part of the programme. The existing clinical data were directed to new patient populations and requirements (for example, inclusion of women of child bearing potential), which was challenged by regulatory authorities at times.
- Approval for use by the company project team (if appropriate) and the therapy area head was required.

When selecting compounds, one of the largest challenges faced was the collation and generation of the relevant datasets, particularly when project teams had been disbanded.



Table 2 | Key differences between the MRC and NIH–NCATS programmes\*

Characteristic	MRC	NIH–NCATS
Purpose	Understanding the mechanisms of disease	Therapeutic development
Scope	Preclinical or clinical or both	Translational projects to hypotheses tested in full Phase IIa trials
Proposal review	AstraZeneca is involved in reviewing the concept proposals	The NIH performs peer-review of concept proposals without company involvement
Industry participants	AstraZeneca only	Eight pharmaceutical companies (AstraZeneca, Pfizer, Eli Lilly & Co., Abbott, Bristol-Myers Squibb, GlaxoSmithKline, Johnson & Johnson and Sanofi)
Time from call for proposals to dosing the first subject in the first study	25 months (sequential process)	14 months (more parallel activities)
Collaborative agreement	Negotiated after MRC project approval using a mCRA template	A signed agreement is required before the final proposal submission. Company template CRA agreements posted online
Funding	3-year award provided up front	Milestone-driven, first year provided up front
Milestone management	Extendable	Preclinical milestone is ≤12 months after funding is awarded; total time allocated ≤36 months
Supervision	No required collaboration meetings	NCATS convenes regular investigator–company collaboration meetings

\*There are also several similarities between the two initiatives, including the use of template agreements and of discontinued compounds only. Companies provide the compounds, regulatory documents and advice, and the MRC or NIH provides funding to the investigators; no funding goes to the pharmaceutical companies. CRA, collaborative research agreement; mCRA, model Industry Collaborative Research Agreement; MRC, Medical Research Council; NCATS, National Center for Advancing Translational Sciences; NIH, National Institutes of Health.

For example, final analysis of histopathology data from a rodent carcinogenicity study of a discontinued project resulted in a new reportable finding for one compound. In another example, one discontinued programme was found to still be on partial clinical hold under the original investigational new drug (IND) application, requiring notification to the UK Investigator. In many cases, Investigator's Brochures had not been updated and needed to be. One of the key learnings for AstraZeneca has been that every project should be fully closed out, including noting incomplete datasets, with a view to potential project repositioning or project reactivation. Final project document summaries should be quickly completed by the original team.

The AstraZeneca partnership with the NRPB in Taiwan and the more recent NIH–NCATS 'Round 2' programmes allowed the inclusion of live compounds. This makes the collation of project data much simpler as there is a project team in place, although the dataset is also more actively evolving as ongoing studies read out. In addition, live

compounds bring their own challenges in terms of balancing confidentiality, focus and intellectual property rights to new inventions. Often quoted is the concern that externally sponsored research (be it clinical or non-clinical) may generate data with negative implications for the original programme (for example, a new safety signal). This is best addressed by engagement of the project team and vetting of the risks versus benefits (risk–benefit assessment) of the additional indication to the entire portfolio, not just the single project, to avoid withholding more novel and interesting live compounds.

#### *Funding and programme management.*

Funding for the studies was provided directly from the public funding body to the principal investigator. It should be noted that the 'in-kind' costs incurred by the companies are not trivial, particularly if remanufacture of the active product, preparation of drug product and placebo, completion of study reports or regulatory documents, and/or patient safety coordination are needed.

The processes and timelines for the implementation of the projects varied between the MRC and the NIH–NCATS programmes, which resulted in differences in timelines to the start of studies, particularly for clinical programmes (TABLE 2). This can lead to issues such as expiration of the clinical drug supply. Differences between the programmes included:

- The NIH–NCATS programme required the inclusion of signed collaboration agreements with the full proposal submission, eliminating the post-funding delay caused by negotiating such agreements, whereas there was no such deadline for negotiated agreements for the MRC collaborations. A balance is needed between the speed of processing applications and reducing unnecessary paperwork associated with ultimately unsuccessful applications.
- For the MRC projects, delays were encountered between the approval of the grant and the institution receiving the funds. For example, some investigators needed to hire personnel for the work and the institution would not allow recruiting to begin until funds were received.
- The MRC committed the full funding at the beginning of the grant and was flexible in allowing extensions of the grant timeline in some circumstances, whereas the NIH–NCATS programme was milestone-driven, with the first year of funding being provided up front and subsequent funding being granted based on the study achieving certain milestones.
- The process in the United Kingdom for implementing certain clinical studies was more sequential (that is, various ethics, regulatory and study-implementation approvals were obtained in sequence), whereas investigators in the United States approached certain activities in parallel.
- NIH–NCATS required some supervision and regular meetings to assess the progress of studies towards milestones.

Nevertheless, the first clinical study results for the MRC and AstraZeneca NIH–NCATS programmes were obtained in the first half of 2014; one example is a mechanistic study from the MRC programme that evaluated the effects of a GABA<sub>B</sub> receptor agonist, AZD3355, on capsaicin-induced cough in healthy volunteers, and another, from the NIH–NCATS programme, is a Phase Ib safety study of saracatinib in patients with Alzheimer disease. Both projects progressed to the second phase of their proposals. Most clinical study results from the MRC and NIH–NCATS projects will be obtained in 2015–2016.

**Collaborative working.** There tend to be strict rules governing externally sponsored studies to ensure the avoidance of influence or bias by the pharmaceutical company over the independent investigator. Concerns regarding possible conflicts of interest are particularly acute for marketed products or compounds in Phase III development. These repositioning programmes did not include such late-stage compounds. Indeed, collaboration was essential to bring together the best scientific and clinical expertise to support these approaches, which were aimed at testing novel hypotheses. This led to the creation of a new AstraZeneca policy for collaborative discussions regarding the construction of full proposals for these investigator-sponsored studies. Beyond the scientific, disease and compound properties, others areas of discussion included the trial design, statistical power, patient safety and decision criteria. Nevertheless, because these remain investigator-sponsored studies, the ultimate responsibility and decisions for all aspects of the study lie with the investigator.

**Pharmacovigilance.** One key element of these collaborative studies was to ensure the appropriate ownership and management of patient safety (pharmacovigilance). The responsibility for industry-sponsored studies rests with the company, who maintain one global Investigator's Brochure and a global safety database to support and ensure harmonization with regulatory and ethics safety-reporting requirements. For externally sponsored studies, the investigator takes on the risk–benefit assessment for the study and the regulatory reporting, with additional requirements to report back to the parent company regarding safety.

Regulatory agencies clearly devolve responsibility for individual studies to the investigator, and this raised the potential for lack of centralized coordination for discontinued compounds — a concern if one compound spawned multiple studies in different territories (for example, raising the question of how to effectively share an observation of a suspected unexpected serious adverse reaction (SUSAR) related to the compound in a US-based study with a UK-based investigator working with the same compound in a separate study). Therefore, AstraZeneca developed a system whereby the company retains ownership of the Investigator's Brochure and required annual updates from investigators to keep this as one core document. Study-specific risk–benefit assessments are documented through a cover note to the Investigator

Brochure and/or the protocol and/or the regulatory filing documents (investigational medicinal product dossier (IMPD) or IND documents). The patient safety database is owned by the company and contracted to a third party for maintenance, timely reporting from investigators and dissemination to all sites working with the same compound. Furthermore, safety data owned by the various investigators are made accessible to other investigators. Thus, AstraZeneca provides a central position in the cross-study evaluation of the safety profile and maintains one harmonized and globally available Investigator's Brochure.

#### **Intellectual property and publications.**

Frequently asked questions regarding these partnerships typically involved intellectual property (IP) and publication rights. In general, any existing IP remained with the owning party and options were established for the company to license any new IP generated by the investigator. Independent of IP, a company may wish to license the data generated by the investigator to support further development. For successful studies not further progressed by the compound originator, it is in the best interest of the company and the academic researchers to find a way to advance the programme for the benefit of patients. Regarding publications, the investigators retained the right to publish with standard provisions that provide for the company to review before submission. Both potential issues were avoided through early discussions regarding motivations of each partner when designing the programmes.

#### **Early indicators of success**

Although these novel models have only been running for a short time in the context of drug development, there are early indicators of success and benefits. Highlights include:

- Greater sharing of valuable, and previously closely guarded, proprietary information: both the open publication of compound information, providing a single summary source of previously unavailable information, and the template legal agreements are unprecedented.
- Crowdsourcing works: there was a diverse range of proposals across institutions for indications beyond those initially anticipated that were submitted within compressed timeframes.
- Improved quality of grants: the collaborative discussions during the construction of full proposals brought together academia and industry to produce higher quality proposals, as suggested by the

assessments given by the peer review committees and the 100% success rate of obtaining ethics and regulatory approvals for studies.

- Novel translational research that may not otherwise have been pursued: companies were not pursuing these studies and these studies would not have been possible for investigators to conduct without access to the proprietary compounds and public funds.
- Generation of new intellectual property: to date, at least two patents have been filed by investigators as a direct result of these activities.
- Two clinical studies have been completed, enabling progression to the next phase of the MRC- or NIH-funded projects.
- Spin-off proposals: proposals made by investigators but not funded by the programme still found a way of moving forward. In one case, AstraZeneca elected to fund a proposal on its own merits, and in another case the investigator accessed alternative sources of funding.
- There has also been the very important benefit of forging closer engagement between the academic and commercial sectors, reducing the perceived barriers and misperceptions, and expanding the knowledge base of compound information.

Saracatinib (also known as AZD0530), a deprioritized compound in the AstraZeneca portfolio, is an especially interesting case study for these repositioning programmes. It is a potent, orally bioavailable inhibitor of SRC tyrosine kinase family members, including FYN kinase, with an expansive preclinical and clinical foundation of research, including Phase II studies in a variety of solid tumours. Ultimately, saracatinib failed to sufficiently modify disease progression in these trials. The general and reproductive toxicology of saracatinib was studied in rat and dog models for up to 6 months. In human studies, the safety profile was such that further clinical investigation was possible. Although the focus had been in oncology, it is now recognized that the SRC kinase family is involved in biology across multiple organ systems. Between the MRC and NCATS programmes, five preclinical or clinical projects, and one additional spin-off project, were funded for further investigation across a diverse range of non-oncological diseases (BOX 1). Data were recently published for one project, showing that saracatinib was effective in reducing spatial memory defects and synaptic depletion in a mouse model of Alzheimer disease<sup>9</sup>.

These results support the evaluation of saracatinib in patients with Alzheimer disease, and clinical studies are taking place as part of this programme<sup>10</sup>.

Perhaps the best indicator of early success is the expansion of these programmes. As previously mentioned, AstraZeneca and the NRPB in Taiwan partnered to initiate a similar program, and both the MRC and the NIH have initiated a second round of calls for proposals and funding. The MRC programme has expanded beyond AstraZeneca to include 68 compounds from 7 pharmaceutical companies (AstraZeneca, GlaxoSmithKline, Johnson & Johnson, Eli Lilly & Co., Pfizer, Takeda and UCB). The NIH programme has expanded to include live development compounds and support for paediatric proposals.

Ultimately, the true measure of success will be either positive clinical data that define a new mechanism involved in impacting human disease or definitive negative data that help to disprove a hypothesis and result in a redirection of attention and resources to more promising avenues.

## Outlook

These pioneering government-sponsored drug repositioning initiatives have successfully piloted a new era of open collaboration and innovation between academia and the pharmaceutical industry on translational research. However, these initiatives have only begun to scratch the surface of the potential opportunities to strengthen and expand such efforts.

The number of compounds involved could easily increase by the participation of additional companies and incorporation of more live compounds, particularly those in early development, as is often done in the field of oncology. Additionally, consideration of off-target effects could be included. Disease-centric charity organizations that have focused their strategies on the translation of research into therapies for patients could partner to fund programmes of interest on compounds that have been made publicly available. Some pioneering charities have embraced repositioning as a strategy, including the Michael J. Fox Foundation, which funded their first repositioning-focused programme in 2010, the Leukaemia and Lymphoma Society and Cancer Research UK (CRUK), although the opportunity remains to specifically leverage the compounds posted publicly.

Start-up companies could contribute compounds to expand their efforts without financial dilution. Once a company has

## Box 1 | The evolving science of saracatinib

Saracatinib (also known as AZD0530) is an especially interesting case study for these repositioning programmes. Saracatinib inhibits the SRC kinase family, and although it was developed for oncology indications, it is now recognized that SRC kinases could be important in diseases related to multiple organ systems. Saracatinib is being investigated in six non-oncological indications as a result of the UK Medical Research Council (MRC) and US National Institutes of Health (NIH) programmes (see the figure).

### SRC kinase as a novel analgesic for cancer-induced bone pain

The hypothesis that SRC kinase is a crucial component of cancer-induced bone pain will be tested through preclinical studies in an animal model of bone cancer pain, investigating the effects of inhibiting SRC on pain-related behaviour, spinal cord neuron phosphorylation and signalling, as well as bone resorption, to identify potential analgesic mechanisms of SRC inhibition. A randomized controlled trial of saracatinib will investigate whether it has analgesic effects in cancer patients with bone metastases.

### SRC inhibitors as potential antipsychotics: human testing with psilocybin

Existing data demonstrate that the SRC kinase pathway is directly involved in the observable symptoms associated with the acute administration of hallucinogens that modulate the 5-HT<sub>2A</sub> receptor, such as psilocybin. This clinical study is evaluating the role of SRC kinase in blocking psychosis induced by infusion of psilocybin.

### Efficacy of saracatinib in treatment of chronic otitis media

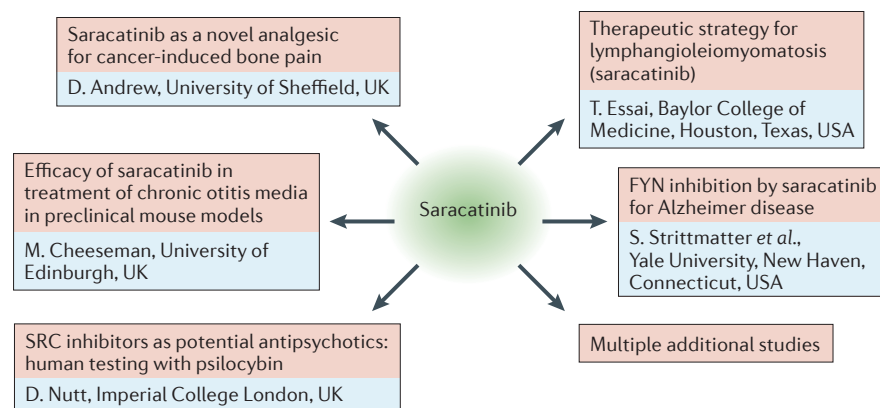
This preclinical study will determine whether local SRC kinase inhibition moderates vascular leak and bulla fluid accumulation leading to reduced hearing loss in models of chronic otitis media.

### Therapeutic strategy for lymphangioleiomyomatosis and tuberous sclerosis

Lymphangioleiomyomatosis (LAM) is a rare progressive cystic lung disease. This research team discovered that SRC kinase is active in LAM cells and is important for cell growth and a cell's ability to move around and invade tissues. This preclinical study aims to determine whether blocking SRC activity is safe and can reduce the growth and the spread of LAM cells.

### FYN inhibition by saracatinib for Alzheimer disease

FYN, a SRC kinase family member, is implicated in triggering Alzheimer disease. This study seeks to test the hypothesis that FYN has an important role in Alzheimer disease and that saracatinib provides benefit to patients with Alzheimer disease in a Phase IIa clinical study. This effort has shown that saracatinib has beneficial effects in a mouse model of Alzheimer disease<sup>9</sup>, and a Phase Ib study of the safety and tolerability of saracatinib in patients with Alzheimer disease has been completed<sup>10</sup>. The drug is to be tested for effectiveness in slowing disease progression in a larger population with mild Alzheimer disease.



participated in one partnership the effort needed for additional partnerships is much lower. Venture capitalists could establish new start-up or spin-out companies based on either the compounds made available or the funded proposals. In fact, one could imagine a 'marketplace' of compounds and funders for investigators to access. As a result of the success of current pilot programmes, AstraZeneca has recently launched a broader

worldwide 'open innovation' initiative that offers a range of compounds for which academic investigators can submit new repositioning ideas and translational research (see Further information). This broad open innovation platform is the first to offer compounds for collaboration and provide a template for a true marketplace that invites ideas and proposals from any contributor and any sector.



A more radical model would use crowdsourcing during the creation of the proposals. At present, proposals are submitted by individuals. Alternatively, an initial proposal, submitted in an open manner, could then be crowdsourced for improvements, including alternative patient populations, endpoints, or trial designs. 'Gamification' — the use of game thinking and mechanics to engage users in solving problems — is one potential crowdsourcing route wherein funders, investigators, companies and reviewers each play a different part. Another possibility is the platform developed by Transparency Life Sciences to develop drugs through collaborative input. Any crowdsourcing model would have to resolve intellectual property rights, determining who performs the research and how credit for the final proposal is assigned.

Barriers remain that could prevent this new partnership model from reaching its full potential. Creating broad awareness of the opportunity among investigators can be challenging. The availability of new compounds should increase over time; however, the available clinical supply of discontinued compounds quickly becomes exhausted. Resistance to the inclusion of live compounds remains prevalent within some companies and funders (although the most recent NIH–NCATS programme did support the inclusion), and the inclusion of biologics is challenged by the complexities of compound supply and delivery. Patent lives are continuously diminishing and the pursuit of more expensive registration studies following positive early studies may be deterred by the lack of financial return. Although regulatory data exclusivity provisions may provide an alternative to patent life, the limited term available in the United States is likely to be insufficient in many cases.

The traditional barriers to open innovation have only just begun to be addressed. Supplemental resourcing, especially funding, is still needed to justify the pursuit of certain areas, including those involving rare and niche indications or those compounds with a limited remaining patent life. If projects are successful, the appropriate distribution of royalties must recognize the relative contributions of each party. Even greater trust

in sharing between parties can be achieved, including sharing scientific information, patient safety monitoring and scientific credit for innovative advances.

Nevertheless, the achievements of these pioneering initiatives include important lessons learned for all parties involved and advances that tackle many of the key barriers. The question confronting all parties now is: will we move forward against any remaining hurdles and advance this model or will we revert to the historical more 'closed' translational research and collaboration models?

Donald E. Frail was previously at the Emerging Innovations Unit, Scientific Partnering & Alliances, AstraZeneca, 35 Gatehouse Drive, Waltham, Massachusetts 02451, USA. Present address: Research and External Science and Innovation, Allergan, Inc., 2525 Dupont Drive, Irvine, California 92612, USA.

Madeline Brady was previously at the Emerging Innovations Unit, Scientific Partnering & Alliances, AstraZeneca, Melbourn Science Park, Cambridge Road, Melbourn, Herts SG8 6EE, UK. Present address: Global Medical Affairs Oncology, Global Medicines Development, Innovative Medicines and Early Development Biotech Unit, AstraZeneca, Melbourn Science Park, Cambridge Road, Melbourn, Herts SG8 6EE, UK.

K. Jane Escott, Alison Holt, and Hitesh J. Sanganeer are at the Emerging Innovations Unit, Scientific Partnering & Alliances, Innovative Medicines and Early Development Biotech Unit, AstraZeneca, Melbourn Science Park, Cambridge Road, Melbourn, Herts SG8 6EE, UK.

Menelas Pangalos is at the Innovative Medicines and Early Development Biotech Unit, AstraZeneca, Melbourn Science Park, Cambridge Road, Melbourn, Herts SG8 6EE, UK.

Chris Watkins is at the Medical Research Council, 14th Floor, One Kemble Street, London WC2B 4AN, UK.

Craig D. Wegner is at the Emerging Innovations Unit, Scientific Partnering & Alliances, AstraZeneca, 35 Gatehouse Drive, Waltham, Massachusetts 02451, USA.

Correspondence to D.E.F.  
e-mail: [frail\\_don@allergan.com](mailto:frail_don@allergan.com)

doi:10.1038/nrd4707

Published online 20 November 2015

1. Ashburn, T. T. & Thor, K. B. Drug repositioning: Identifying and developing new uses for existing drugs. *Nat. Rev. Drug Discov.* **3**, 673–683 (2004).
2. Cook, D. *et al.* Lessons learned from the fate of AstraZeneca's drug pipeline: a five-dimensional framework. *Nat. Rev. Drug Discov.* **13**, 419–431 (2014).
3. Mullard, A. Could pharma open its drug freezers? *Nat. Rev. Drug Discov.* **10**, 399–400 (2011).

4. Carroll, J. Washington U gets second look at Pfizer molecules. *Fierce Biotech* [online], <http://www.fiercebiotech.com/story/washington-u-researchers-get-second-look-pfizer-molecules/2010-05-18#ixzz33uYUlr8> (2010).
5. Collins, F. S. Mining for therapeutic gold. *Nat. Rev. Drug Discov.* **10**, 397 (2011).
6. LaMattina, J. The NIH is going to discover drugs... Really? *Forbes* [online], <http://www.forbes.com/sites/johnlamattina/2012/05/15/the-nih-is-going-to-discover-drugs-really/> (2012).
7. Colvis, C. M. & Austin, C. P. The NIH-industry New Therapeutic Uses pilot program: demonstrating the power of crowdsourcing. *Drug Repurp. Rescue Repos.* **1**, 15–16 (2015).
8. Southan, C., Williams, A. J. & Ekins, S. Challenges and recommendations for obtaining chemical structures of industry-provided repurposing candidates. *Drug Discov. Today* **18**, 58–70 (2013).
9. Kaufman, A. C. *et al.* Fyn inhibition rescues established memory and synapse loss in Alzheimer mice. *Ann. Neurol.* **77**, 953–971 (2015).
10. Strittmatter, S. M. Safety and tolerability of AZD0530 (Saracatinib) in Alzheimer's disease. *ClinicalTrials.gov* [online], <https://clinicaltrials.gov/ct2/show/NCT01864655?term=strittmatter&rank=1> (2014).

#### Acknowledgements

The UK Medical Research Council (MRC) and the US National Institute of Health (NIH) initiatives were supported by a large range of individuals within the MRC and the NIH as well as external investigators, and investigators from AstraZeneca and MedImmune. At the risk of excluding many, in particular the authors would like to thank J. Latimer (MRC), C. Colvis and B. Dunn (NIH–NCATS), G. Wilkinson, C. Wilks, A. Longton, S. Curran, K. Hickling and the members of the New Opportunities Emerging Innovations Unit (AstraZeneca). The authors would like to acknowledge the contribution made by all those involved and the exciting ideas and proposals received from academic investigators.

#### Competing interests statement

The authors declare **competing interests**: see Web version for details.

#### FURTHER INFORMATION

AstraZeneca Open Innovation: <http://openinnovation.astrazeneca.com/>  
MRC Translational Research Strategy: <http://www.mrc.ac.uk/research/initiatives/experimental-medicine>  
Nature focus on challenges in reproducible research: <http://www.nature.com/nature/focus/reproducibility/>  
NCATS Round 2 Industry provided agents (2014): <http://www.ncats.nih.gov/research/reengineering/rescue-repurpose/therapeutic-uses/directory2014.html>  
NCATS Pilot Programme (2012): <http://www.ncats.nih.gov/research/reengineering/rescue-repurpose/therapeutic-uses/directory.html>  
NIHR/MRC model Industry Collaborative Research Agreements: <http://www.ncrri.nihr.ac.uk/resources/micra/>  
NIH–NCATS: Discovering New Therapeutic Uses for Existing Molecules. 2012 URL: <http://www.ncats.nih.gov/research/reengineering/rescue-repurpose/therapeutic-uses/therapeutic-uses.html>  
Southan figshare page: [http://figshare.com/articles/NCATS\\_Compounds\\_with\\_identifications/92850](http://figshare.com/articles/NCATS_Compounds_with_identifications/92850)  
MRC-Industry Asset Sharing Initiative (2014): <http://www.mrc.ac.uk/news-events/news/world-s-largest-collection-of-deprioritised-pharma-compounds-opens-to-researchers/>  
UK Government Lambert agreements: <https://www.gov.uk/model-agreements-for-collaborative-research>

#### SUPPLEMENTARY INFORMATION

See online article: S1 (table) | S2 (table)

ALL LINKS ARE ACTIVE IN THE ONLINE PDF

## Towards a hit for every target

Steve Rees, Philip Gribbon, Karen Birmingham, William P. Janzen and Garry Pairaudeau

Technological advances coupled with novel collaborative strategies for compound sourcing and management are poised to transform the utility of high-throughput screening.

The advent of high-throughput screening (HTS) in the 1990s revolutionized the search for lead molecules. It enabled screening of large, diverse compound libraries against drug targets to identify chemical probes for target validation and drug candidates for clinical development on a scale not previously envisaged. However, the success of HTS was limited by the quality and diversity of the screening libraries and by the nature of many of the assay technologies, which required compound testing in artificial biological systems<sup>1</sup>.

Twenty-five years on, we believe we are on the cusp of a new era in HTS, in which its transformation from a 'numbers game' to a smart selection tool will be completed. This transformation is being driven by recent advances in computational and medicinal chemistry, cell biology, biophysics and instrumentation, which together could enable the design of bespoke screening libraries for each drug target that can be tested in information-rich biophysical assays and advanced cellular models. The innovations now being implemented could deliver on the long-sought promise of a hit for every target, meeting the strategic need to look beyond traditional drug target families for the next generation of medicines. Furthermore, a range of innovative partnerships mean that in this new era, lead discovery based on HTS will no longer be the preserve of pharmaceutical companies, but will become commonplace in the academic research community.

### Enhanced screening technology

For much of the past 20 years, pharmaceutical companies have used a target-class approach to drug discovery, focused on screening members of tractable protein superfamilies such as G protein-coupled receptors and kinases. A major strategy was to identify chemical probes to be used for associating the targets with disease. There is now a drive towards the selection of drug targets with increased levels of target validation before a screening programme is started<sup>2</sup>. This has led to an increase in the diversity of target types and biological pathways to be screened, and includes a renewed interest in phenotypic screening.

However, advances in assay biology and screening technology have been slow. Testing compounds against recombinant proteins or cell lines in fluorescent, luminometric or radiometric assays in microtitre plate format

has been the workhorse of the field. We believe this situation will evolve. Stem-cell and primary-cell technology, together with 3D and microphysiological culture systems, will increasingly enable screening in more disease-relevant assays. The use of 'omics' platforms (transcriptomic, proteomic and metabolomic), in which multiparametric data from complex assay systems are collected, will become commonplace, and advanced data capture and analytics tools will be used to interpret the large volumes of data. Finally, new technologies, including acoustic delivery of samples for mass spectrometry, will enable novel screening approaches.

### Improved compound library quality

A typical pharmaceutical company compound collection consists of ~1–3 million molecules selected for their drug-like or lead-like properties<sup>3</sup>. These collections represent the chemical history of companies based on their historical portfolio of internal projects, directed efforts to target libraries to specific areas of biological relevance, and synthesis or acquisition based on chemical diversity and synthetic tractability.

Success in screening against novel target families requires that we expand into new chemical space and new therapeutic modalities, yet the cost of acquiring or making large numbers of new compounds is prohibitive even for large pharmaceutical companies. Moreover, practical considerations such as storage volume, store size and screening costs limit the number of compounds in a collection. We believe that these constraints can be addressed through a series of imminent innovations.

First, knowledge-based selection of compounds for screening is limited to target classes for which knowledge (typically structure-derived) of desirable pharmacophores is available, such as kinases, and the size of such sets is limited by the capacity of available compound-picking robots. In the coming years, technologies such as cryo-electron microscopy and advances in the ability to crystallize membrane proteins should rapidly expand knowledge of the structure of putative drug binding sites. Combined with improved computational tools that enable more effective interrogation of chemical space and the development of capability in compound management to create bespoke screening libraries, it will become possible to create knowledge-based screening libraries

Steve Rees is vice president, Screening Sciences and Sample Management at AstraZeneca, Cambridge Science Park, 310 Milton Road, Cambridge CB4 0FZ, UK. Garry Pairaudeau is head of External Sciences at AstraZeneca, Cambridge Science Park, 310 Milton Road, Cambridge CB4 0FZ, UK. Philip Gribbon is coordinator of EU-OPENSREEN and at the Fraunhofer Institute for Molecular Biology and Applied Ecology Ecology IME ScreeningPort, Schnackenburgallee 114, 22525 Hamburg, Germany. Karen Birmingham is science media relations director, AstraZeneca, 2 Kingdom Street, London W2 6BD, UK. William P. Janzen is executive director, Lead Discovery, Epizyme, 400 Technology Square, Cambridge, Massachusetts 02139, USA.

Correspondence to S.R.  
[Steve.Rees@astrazeneca.com](mailto:Steve.Rees@astrazeneca.com)

doi:10.1038/nrd.2015.19  
 Published online 20 Nov 2016

for many targets. Delivery of this vision will require the development of new high-throughput synthetic methods to create compounds for screening.

Second, the ability to dispense nanolitre volumes of a compound solution directly from low-volume storage tubes to assay plates using acoustic technology will revolutionize compound library design, storage and distribution through miniaturization. It will enable the screening of compounds generated by expensive or low-yield syntheses and will reduce storage costs. The ability to rapidly pick compounds will support the creation of target-specific compound libraries and iterative screening paradigms. This could have a particular impact in the academic setting, as publicly sourced compounds are typically available in small amounts.

Third, compound sharing creates the opportunity to rapidly increase library diversity. For example, AstraZeneca has recently made agreements with Sanofi and Syngenta to exchange 240,000 novel compounds for screening at no cost. Many examples of sharing exist with and between academic centres, including Sanofi and Evotec, Sanofi and Fraunhofer's Center of Excellence for Natural Products Research, and the University of North Carolina's agreements with Southern Research Institute, the National Center for Advancing Translational Sciences and the LIMR Chemical Genomics Center. Another model of compound sharing is the collaboration between AstraZeneca and Bayer, whereby each company makes available its entire compound library for screening of the partner's target.

Finally an increased interest in screening biological molecules, including peptides and secretome-, siRNA- and CRISPR-based libraries will require the development of new methods for sample storage and supply to complement small-molecule screening.

Only the largest organizations could create these capabilities in-house. The smartest will create networks of shared capabilities to match the right drug target to the right compounds.

### Succeeding through partnership

There has been a recent and dramatic increase in the number of academic drug discovery centres; the [Academic Drug Discovery Consortium](#) currently includes 136 centres. This increase has been driven by growing interest in identifying small-molecule probes to explore biology and in therapeutic discovery, as well as a strategic intent from funders in translational science.

In parallel with the changes in academia, there has been a shift in the mindset of some companies to embrace 'open innovation'<sup>24</sup>, which recognizes that no organization can succeed alone and partnerships based on complementary investment and shared success are needed to deliver novel medicines. In hit discovery, this requires companies to make their libraries and screening infrastructure available to partners, and for academics to share target and disease knowledge and expertise across multiple institutions, including companies.

Some examples of this paradigm already exist. For example, the creation of the European Lead Factory (ELF) involved the donation of compounds from seven

companies to create the 270,000-compound Joint European Compound Library (JECL), which is being supplemented by 230,000 novel academic compounds. The EU-OPENSOURCE project aspires to create a network of specialist laboratories sharing a common infrastructure and compound library, where users will be mandated to make project bioactivity results available to the broader scientific community.

Several companies are making all or part of their compound library available to academia through open innovation collaborations. For example, AstraZeneca's MRC Centre for Lead Discovery will see scientists from the Medical Research Council (MRC), Cancer Research UK (CRUK) and AstraZeneca working side by side to advance industrial and academic targets using next-generation compound management and screening platforms. AstraZeneca will make its hardware, compound library and scientific expertise available to academia; the MRC and CRUK will identify innovative drug targets from their principal investigator networks.

### Training and skills

The skills and knowledge in compound management and HTS that were traditionally the preserve of large pharmaceutical companies and contract research organizations are increasingly needed in smaller companies and academia. Until now, such expertise has largely come from the release of experienced staff from industry who have moved into academia. However, the explosion in the number of academic drug discovery groups is creating a skills shortage at the same time as generating a solution. Many young scientists are completing their formal training with experience in these areas gained through collaborations with academic drug discovery centres. The remaining gap is the formalization of these training programmes, which must be addressed by funding bodies. Increasingly, organizations such as the [Society for Laboratory Automation and Screening](#) are facilitating professional programmes to train and develop the coming generation of hit discovery scientists.

### Conclusion

Attaining the holy grail of a hit for every target depends on new ways of working and on technological and scientific advances. Industry and academia must invest in their respective strengths of lead and target identification and both must open their doors and their expertise to partnerships based on complementary investment and shared success. Funding bodies should foster such collaboration and companies should create innovative win-win deal structures to encourage such partnerships.

1. Macarron, R. *et al.* Impact of high throughput screening in biomedical research. *Nat. Rev. Drug. Discov.* **10**, 188–195 (2011).
2. Cook, D. *et al.* Lessons learned from the fate of AstraZeneca's drug pipeline: a five-dimensional framework. *Nat. Rev. Drug Discov.* **13**, 419–431 (2014).
3. Bakken, G. A. *et al.* Shaping a screening file for maximal lead discovery efficiency and effectiveness: elimination of molecular redundancy. *J. Chem. Inf. Model.* **52**, 2937–2949 (2012).
4. Simpson, P. B. & Reichman, M. Opening the lead generation toolbox. *Nat. Rev. Drug Discov.* **13**, 3–4 (2014).

### Competing interests statement

The authors declare [competing interests](#): see Web version for details.



# The global intellectual property landscape of induced pluripotent stem cell technologies

MacKenna Roberts<sup>1</sup>, Ivan B Wall<sup>1–4</sup>, Ian Bingham<sup>5</sup>, Dominic Icelly<sup>5</sup>, Brock Reeve<sup>6</sup>, Kim Bure<sup>7</sup>, Anna French<sup>1,10</sup> & David A Brindley<sup>1,6,8–10</sup>

**Will freedom to research and innovate be restricted as the induced pluripotent stem cell field advances toward the clinic, or are concerns premature within a rapidly changing ecosystem?**

Intellectual property (IP) rights lie at the core of the commercialization process, serving as a powerful incentive to harness the potential of technologies for therapeutic applications. However, when filed inappropriately—with broad or premature claims, for example—or when mismanaged, patents can obstruct vital precompetitive collaborations, dampen investor interest and threaten clinical translation and patient access<sup>1–3</sup>. Moreover, uncertainties, including those around the validity of claims or the priority of ownership rights, can depress progress<sup>4</sup>.

Stem cell derivation and cellular reprogramming techniques are broadly

enabling technologies for which ready access is essential to a sustainable cell therapy industry. Pluripotent stem cells (PSCs) have the potential to transform future healthcare. Clinical trials using embryonic stem cells (ESCs) have begun to show early promise<sup>5</sup>; however, concerns remain regarding limited supply and ethics. In 2007, scientists at Kyoto University and, independently, at the University of Wisconsin, reprogrammed adult human skin cells to generate induced pluripotent stem cells (iPSCs)<sup>6,7</sup>. In addition to long-term therapeutic potential, iPSC technologies are valuable tools for drug discovery, toxicology testing and disease modeling.

Although the total number of stem cell patent filings has declined since 2008, patents for iPSC technologies continue to increase<sup>8</sup>. Analysis of the blistering pace of scientific progress in this area indicates that the growth is unlikely to abate. Widespread concern has been voiced that the emerging ecosystem is becoming burdened by prohibitive and cumulative licensing fees that could restrict scientists' freedom to research and patients' equitable access to resulting medical benefits<sup>2–4,8,9</sup>. Although these concerns may be allayed through innovation and industry growth cycles, the upstream production and downstream differentiation of iPSCs into desired cell lineages for application requires numerous interrelated, complex technologies. This distinguishes cellular patents from the evolution of other highly patented industries such as small-molecule drugs and electronics. No single company currently controls the IP for all techniques, methods and reagents required for the production of iPSCs<sup>10</sup>. A global race is underway to establish the most suitable and efficient methods for each of these component technologies.

Although it may be beneficial for competing industry that no single entity dominates market access to all fundamental iPSC technologies, congestion of patents related to broadly enabling technologies, referred to as a 'patent thicket', has the potential to bottleneck innovation. If ownership of overlapping fundamental aspects is spread across several patent holders, research may become encumbered by the negotiation of complicated stacked licensing with prohibitive fees and royalty schemes that drain value from innovation. This is referred to in IP rights theory as 'anti-commons' innovation<sup>11</sup>. Funders may, consequently, be deterred by vulnerability to future challenges against IP ownership of a product after costly development.

Analysts estimate that total global revenues for companies that supply iPSC research products was \$837 million in 2012 and that this will exceed \$1 billion by 2015 (ref. 12). Additionally, 22% of stem cell researchers report having used iPSCs<sup>12</sup>. With first-in-man clinical research studies underway this year at RIKEN<sup>4</sup>, now is a critical time in iPSC technology R&D to address the validity of any potential IP obstacles to the commercial translation of medical advancements.

Despite extensive literature, a current, global analysis that focuses exclusively on the burgeoning IP landscape for iPSC technologies has not been performed. To provide clarity, we consider the broad landscape and discuss the validity of mounting concerns over licensing practices for this nascent ecosystem, possible solutions and the emergence of other forms of IP that could erect barriers to open and collaborative science. Accordingly, we have employed commercial software to analyze patent

<sup>1</sup>Oxford–University College London Centre for the Advancement of Sustainable Medical Innovation, University of Oxford, Oxford, UK.

<sup>2</sup>Department of Biochemical Engineering, University College London, London, UK.

<sup>3</sup>Department of Nanobiomedical Science and BK21 Plus NBM, Global Research Center of Regenerative Medicine, Dankook University, Cheonan, Republic of Korea. <sup>4</sup>Biomaterials and Tissue Engineering Lab, Department of Nanobiomedical Science and World Class University Research Center, Dankook University, Cheonan, Republic of Korea. <sup>5</sup>IP Asset LLP, Oxford, UK. <sup>6</sup>The Harvard Stem Cell Institute, Cambridge, Massachusetts, USA. <sup>7</sup>Sartorius Stedim, Göttingen, Germany.

<sup>8</sup>Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences, Nuffield Orthopaedic Centre, University of Oxford, Oxford, UK. <sup>9</sup>Centre for Behavioural Medicine, UCL School of Pharmacy, University College London, London, UK. <sup>10</sup>These authors contributed equally to this work.

e-mail: [anna.french@csmi.org.uk](mailto:anna.french@csmi.org.uk) or [david.brindley@ndorms.ox.ac.uk](mailto:david.brindley@ndorms.ox.ac.uk)

**Table 1 General iPSC technology category distribution of estimated patent filings: production (upstream technologies) versus differentiation (downstream technologies)**

	iPSC main technology categories	
	Production	Differentiation
Estimated no. of patent families (patented inventions)	650	1,300–1,400

data, which identified a fragmented patent landscape with internationally dispersed ownership around complex, interrelated technologies.

### Technology areas and trends

To identify the major trends, a raw data set was harvested to include a broad range of technologies related to iPSCs, comprising 4,651 documents with 1,388 Derwent World Patents Index (DWPI) families (patented inventions) using Thomson Innovation software. This initial collection was characterized and cleaned, both manually and using Thomson Data Analyzer software (see **Supplementary Note**). Patents filed and/or granted between 1 September 2006 and 31 December 2013 were searched using keywords in patent titles, abstracts and claim sections across a complete global data set of 50 patent authorities.

The challenge in isolating the iPSC patent landscape from the broader pluripotent landscape (**Supplementary Fig. 1**) is in the high degree of overlap in terminology, techniques and applications that relate to both ESCs and iPSCs. Another particular challenge is the significant overlap between cellular processes and materials claimed for reprogramming and differentiation.

To narrow the search within pluripotency, we selected more rigorous keyword searches (**Supplementary Table 1**) to yield data sets that were further manually refined through random sampling of more than 700 patents (**Supplementary Table 2**). The data set therefore represents an over-inclusive estimation rather than a precise figure.

The iPSC R&D pipeline is divided between upstream production and downstream differentiation techniques. Production refers to the reprogramming of somatic cells to a pluripotent state. Once reprogrammed, these cells are selected and replicated. Candidate pluripotent cell colonies are then characterized for confirmation of a pluripotent phenotype before further propagation. The resulting iPSCs can serve as important research tools for a range of downstream applications that differentiate iPSCs into specialized cells. The differentiated cells created from PSCs can then be used for drug screening, toxicity testing, disease modeling or therapeutic application.

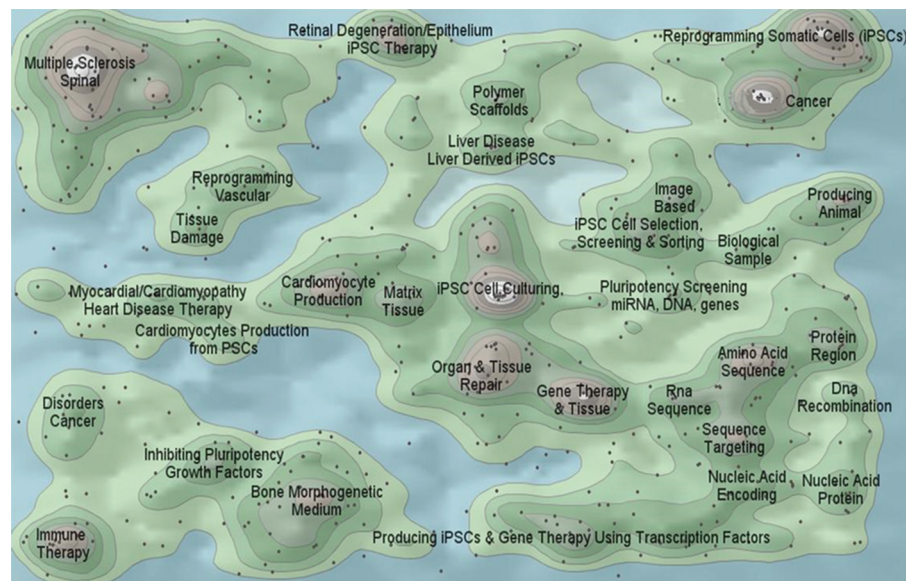
Differentiation technologies are the largest group of iPSC patents. In this broad category of patents, referred to as the 'differentiation patents', iPSCs have a range of roles, from central to more ancillary. In the latter, pluripotent cells are cited merely as a possible starting material for the invention, among other listed cell types. Including differentiation patents within the iPSC landscape initially gives a broad overview with considerable overlap with the wider PSC (iPSC and ESC) landscape. A conservative estimate of the number of differentiation patents is 1,300–1,400 patent families (inventions) with 4,500–5,000 documents overall (**Table 1** and **Supplementary Table 3**). This figure would probably be higher if more differentiation patents for ancillary technologies that cite iPSCs in their claims were included.

The category of patents related to iPSC production, an estimated 650 filings, includes culture techniques, growth factors and

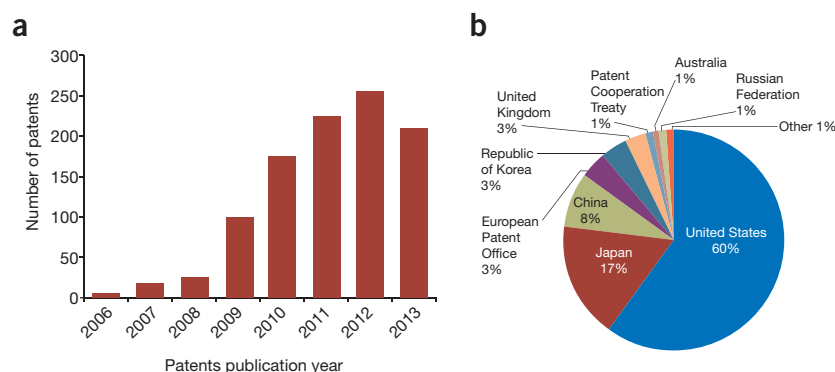
technologies essential to the generation of iPSCs (**Table 1** and **Supplementary Table 3**). Estimates do not account for discontinued or denied applications. Excluding ancillary technologies such as culture techniques, we identified 299 patents for reprogramming methods.

A digital representation of the broad iPSC patent landscape provides a macroscopic glimpse of patenting trends and can help visualize technology clusters (**Fig. 1**). Automated software clusters similar terminology found in the patents' titles, claims and abstracts. Patents are represented by dots, and those with similar terminology are delineated by contour lines. The greater the similarity in the patents' terminology, the greater the elevation of the cluster contour. The white 'snow-capped' peaks represent areas of high patent activity for the labelled terms: multiple sclerosis, spinal injury and diabetes research; cancer research; reprogramming methodologies; gene therapy and tissue engineering; and cell culturing mediums and techniques, including assays to confirm pluripotency and cell population selection techniques (**Fig. 1**).

Although a number of the peaks involve cell-reprogramming techniques distributed among a range of technology areas



**Figure 1** Broad iPSC technology patent landscape. The five white 'snow-capped' peaks indicate technology areas of high patent activity: cell culture (including cell selection and characterization techniques), spinal injury, cancer (involving cancer cells, inhibiting teratoma formation or research intended as a therapeutic indication toward cancer), reprogramming methodologies and patent documents that cite gene therapy or tissue engineering purposes. Each dot represents a patent. The stronger the correlation of automatically clustered terminology within a contour, the higher the elevation of related patents. The patent map, generated using Thomson Innovation software, provides a representative overview of the original raw patent collection before it was cleaned to remove patents that do not directly claim iPSCs as a central aspect of the patented invention. This initial raw data set of 1,388 patent families with 4,651 total documents comprises both granted and pending applications from 1 September 2006 to 31 December 2013. (Source: Thomson Reuters)



**Figure 2** Trends in iPSC patent document filings. **(a)** The number of total patented inventions filed for iPSC technologies has increased each year since 2006, peaking in 2012. **(b)** National contributions to iPSC invention by earliest priority country. The United States dominates the fundamental technology platforms in the field, followed by Japan, as could be predicted on the basis of technology origin and subsequent national focus. China and the Republic of Korea significantly contribute in addition to the European Patent Office and the United Kingdom.

(Supplementary Fig. 2), the map flushes out a visual of the third-largest category of iPSC-related patents (generalized from the size of the snow-capped peaks in Fig. 1): cell-culturing techniques and growth factors. This category includes technologies for both the generation and differentiation processes.

Each of these broad categories represents a technology umbrella for a series of complex, multistep processes. Therefore, each step in the iPSC 'pipeline' further subdivides into specific areas of research and IP proliferation. Because accepted best-practice standards within the industry for each of these steps do not yet exist, they represent current

and potentially profitable technology gaps but, equally, concerns for patent thickening.

#### Patent filing profile

Published patent applications filed on iPSC-related technologies nearly doubled from 2009 to 2010 (Fig. 2). In 2012, Shinya Yamanaka shared the Nobel Prize in Physiology or Medicine with John Gurdon for their contributions to reprogramming. In the same year, published patent applications relating to iPSCs reached an all-time high (Fig. 2). Despite a modest dip in 2013, a similar number (150) of new filings on

discoveries broadly relevant to iPSCs can be expected in 2014. This trend is unlikely to decrease, provided a nonobvious, novel step change in reprogramming standards is not discovered.

Patent activity was analyzed by priority country to indicate geographical concentration of iPSC research. The United States and Japan clearly lead, with the greatest contributions to iPSC patented inventions, but research is active on a global stage (Fig. 2). In 2013, Japan passed legislation to expedite approval of candidate stem cell therapeutics without phase 3 clinical trials<sup>13</sup>. If this regulation results, as anticipated, in the acceleration of therapeutic products to market, the corresponding patents may also be expedited, as was the grant of Yamanaka's patent in 2008. However, this remains to be seen. Meanwhile, the United States has steered in a different direction. The federal appeals court ruled in February 2014 that more-than-minimally manipulated stem cell therapies must be regulated as a (biological) drug in accordance with US Food and Drug Administration regulations, rather than more leniently as a medical practice or medical device (cellular product)<sup>14</sup>.

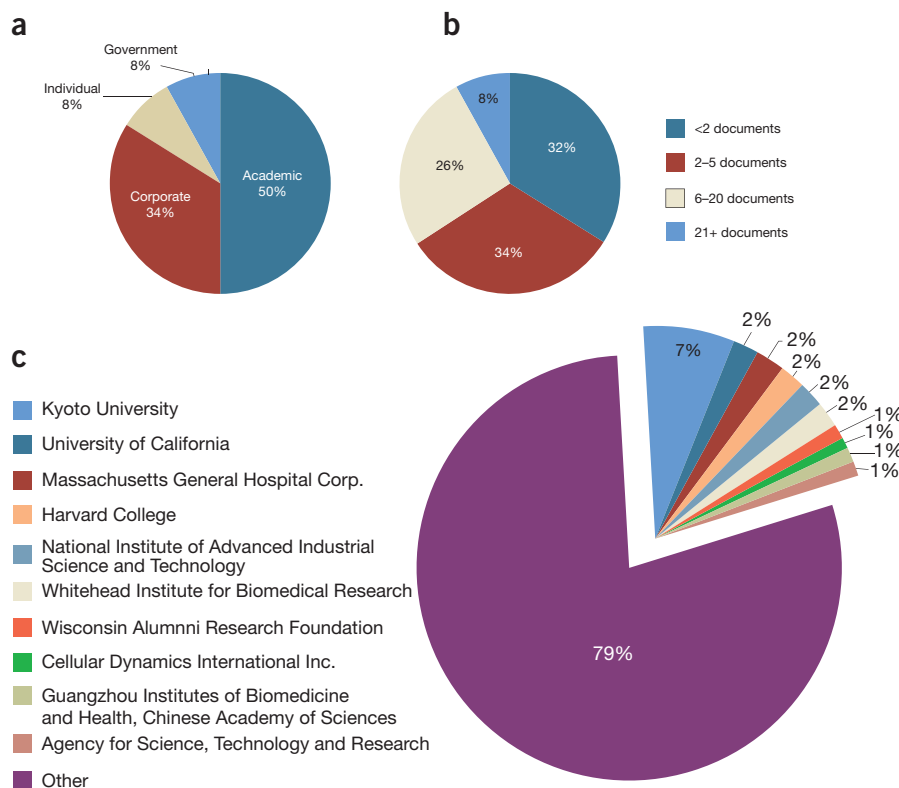
Still, the United States, with its significant margin for pending and issued patents, is unlikely to be surpassed by Japan. Additionally, the United States has a more diverse base of assignee expertise compared to Japan's concentration in a handful of institutions (Table 2). However, in an

**Table 2** Leading 10 assignees and inventors (924 patent family cleaned subset of collection)

Rank	Organization	Top inventors
1	Kyoto University (Kyoto, Japan)	Yamanaka, S., Takahashi, K., Nakagawa, M.
2	University of California, Santa Cruz (Santa Cruz, California)	NA
2	General Hospital Corp. (Boston)	Hochedlinger, K., Maherali, N.
2	Harvard College (Boston)	Maherali, N., Hochedlinger, K.
4	National Institute of Advanced Industrial Science and Technology (Kyoto, Japan)	Yamanaka, S., Goshima, N., Mochizuki, H., Kawamura, Y., Maekawa, M.
4	Whitehead Institute of Biomedical Research (Cambridge, Massachusetts)	Jaenisch, R., Hochedlinger, K., Young, R.A.
4	Wisconsin Alumni Research Foundation (Madison, Wisconsin)	Thomson, J.A., Chen, G., Yu, J.
5	Cellular Dynamics International (Madison, Wisconsin)	Mack, A., Yu, J., Thomson, J.A.
5	Guangzhou Institutes of Biomedicine and Health (Guangzhou, China)	Pei, D.Q., Esteban, M., Quin, D.J., Chen, J.K., Yang, J.Y.
6	Agency for Science, Technology & Research (Singapore)	Ng, H.H.
6	Salk Institute for Biological Studies (La Jolla, California)	Belmonte, J.C.I., Gage, F.H., Raya, A.
6	Stanford University (Stanford, California)	NA
7	Advanced Cell Technology (Marlborough, Massachusetts)	Lanza, R., West, M.C., Lu, S.J.
7	Sangamo Biosciences (Richmond, California)	Holmes, M.C., Gregory, P.D., Urnov, F.
7	University of Tokyo (Tokyo)	Nakauchi, H., Takayama, N., Eto, K.
8	Johns Hopkins University (Baltimore, Maryland)	NA
9	Keio University (Tokyo)	Okano, H., Fukuda, K., Hattori, F.
10	Children's Medical Center (Boston)	Daley, G.Q., Agarwal, S., Park, I.H.
10	Primegen Biotech (Irvine, California)	Silva, F.J., Sayre, C.B.

Only inventors with three or more patents are named. NA, not applicable.





**Figure 3** The distribution of ownership origination shows a fragmented landscape with public-sector dominance and a growing cottage biotech industry. **(a)** Breakdown of sector participation in iPSC R&D shows a significant number of iPSC patents in government, nonprofit and academic institutions (58%). The public sector dominance is unsurprising for emerging technologies. However, 34% corporate ownership indicates that this field has started to actively privatize at an early stage. **(b)** Distribution of patented inventions among patent holder organizations. The domination of lower-value patent portfolio holdings shows a large number of smaller players in the sector, indicating that the 34% corporate face of iPSC patent ownership is composed largely of small biotechnology companies. The thin distribution of patent ownership reveals a highly fragmented landscape. **(c)** The ten leading organizations, determined on the basis of overall patent-document numbers for iPSC inventions, shows Kyoto University as the top iPSC patent holder (both for overall applications and granted patents). Despite the overall dominance of the United States, Japan has two institutions, Kyoto University and the National Institute of Advanced Industrial Science and Technology, in the top ten organizations for iPSC inventions, in addition to the Guangzhou Institutes of Biomedicine and Health in China and the Agency for Science, Technology and Research in Singapore.

assessment of patents' value to the field through citation analysis compared with the location of the top assignees (patent holders), the geographical hubs for iPSC research are Japan, Massachusetts, California, Wisconsin and the United Kingdom (Table 2, Table 3 and Fig. 2).

China and South Korea both have strong footing in the top five countries filing iPSC patents (Table 2 and Fig. 2). Their positioning is stronger for iPSCs than for PSCs or general stem cell patents<sup>15</sup>. In terms of patent filing numbers, China appears to be the third strongest contributor to iPSC-related inventions exceeding the European Union (EU). The EU has a similar patent activity level to South Korea (Fig. 2). This distinguishes the iPSC patent landscape. By contrast, the EU is the third highest contributor with regard to

patents filed for embryonic stem cells discovery and stem cell research more generally. Although these geographical comparisons include more ancillary iPSC technologies and do not account for validity of the patents filed, they clearly indicate greater global participation than other areas of stem cell research<sup>15</sup>.

After the United Kingdom, Germany is the top patent-filing country in the EU for all stem cells<sup>15</sup>. For the iPSC patent landscape, the United Kingdom remains the strongest patenting EU member state. Germany is active in the PSC space but does not seem currently to have a leading presence in iPSC technology (Fig. 2). German institutions, however, are the second largest contributors, behind France, of regional EU patent applications. It is possible that a stronger presence has yet to emerge owing to the year-and-a-half time from filing to publication.

Even in the broader pluripotent cell landscape, Kyoto University is the established leader among patent applicants<sup>15</sup>. However, until as recently as 2012, the Wisconsin Alumni Research Foundation was the institution granted the most stem cell patents<sup>15</sup>. Unsurprisingly, in the iPSC space, Kyoto University leads for total filed and granted patent applications. Unlike the majority corporate ownership of stem cell technologies<sup>15</sup>, iPSC technologies are emerging and, therefore, still dominated by public sector R&D, with only 34% of patents owned by corporate entities (Fig. 3a).

### Patent distribution and patent thickets

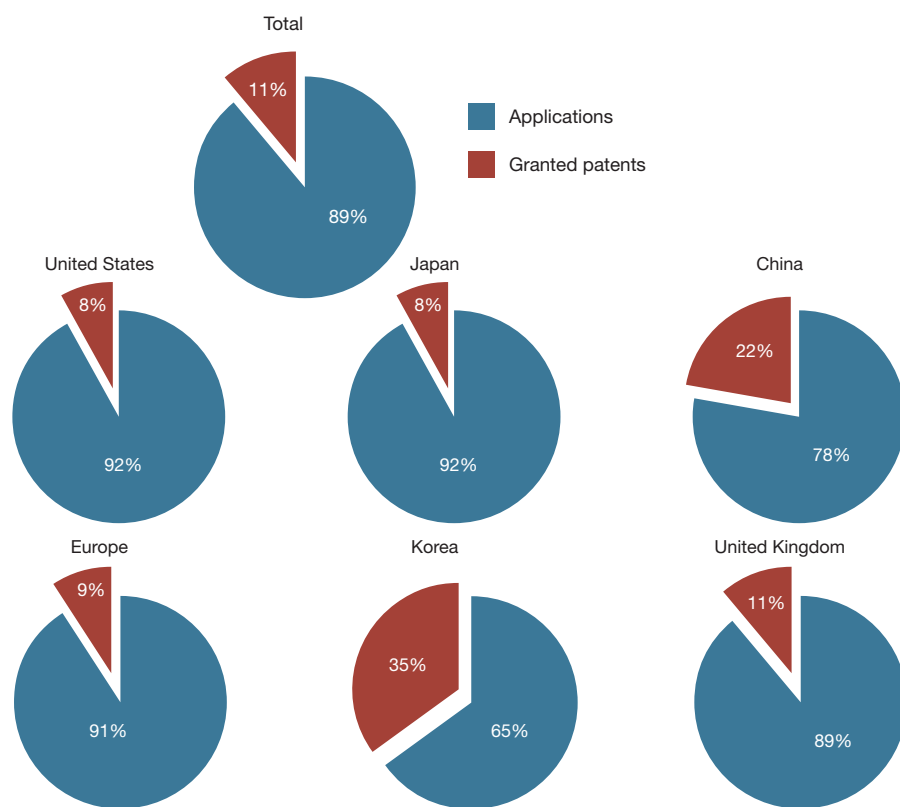
Patent portfolios remain small and thinly distributed among many entities: 66% of applicants own fewer than five patent applications, and 32% have one patent application (Fig. 3b). IP dominance over iPSC technologies is established by relatively small holdings of six to more than 21 filings in a portfolio (Fig. 3b). The total number of patents owned may be higher than shown, because an assignee may hold several patents worldwide for any single invention. Here, we restricted analysis to the number of inventions to determine an entity's impact as well as portfolio diversity.

The ten leading assignees and inventors at leading institutions (Fig. 3c) reveal a significant overall international contribution to the field, indicating a global dispersion of basic iPSC technologies.

The growing disparity between granted and pending applications further clouds the certainty of what parties may validly claim (and license) which technologies. Regulatory and legal challenges have contributed to an increased delay in examination, particularly before the European Patent Office, although granted patents are still subject to challenge. Only 11% of patents are issued, and an overwhelming 89% of applications await examination (Fig. 4).

This distribution pattern is typical of an emerging technology. However, thin patent distribution among industry also warrants concern for patent-thicket formation. Strong competition by many actors in the field can fuel developments, but it also indicates that the task of coordinating access to interrelated enabling technologies might require an onerous and complex licensing scheme.

Despite the public sector's dominance, a substantial number of smaller biotech companies occupy the other end of the filing spectrum. Many of these have filed only one iPSC-related patent. The 34% contribution from the private sector therefore constitutes



**Figure 4** The percentage of patents granted globally is low and varies between countries. Only 11% of all filed iPSC-related patents have been granted, leaving 89% pending examination, adding to the uncertainty of who will hold rights in what technologies. The ratio of iPSC granted patents to pending applications for each of the leading five patenting nations and the EU (through its regional patent application process) by priority country is shown.

a large number of companies holding a low number of patents (Fig. 3a). Further in-depth analysis of company profiles and collaborations would probably prove to be of interest.

This analysis does not reflect the licensing of portfolios to 'exploitation companies' that privately manage the licensing arrangements for a range of patent holders. By way of example, iPS Academia Japan (iPS AJ) is a subsidiary of Kyoto University that is privately funded and has widely sublicensed iPSC patent portfolios from several universities and research institutions in Japan and beyond, including Axiogenesis AG of Cologne, Germany, and Auckland UniServices Ltd. of New Zealand. Its total licensed patent portfolio includes an estimated 74 families (an estimated 280 applications with 70 patents granted worldwide). iPS AJ manages IP for iPSC-related technologies, cellular products, drug discovery and cell therapy studies. All licenses are nonexclusive; those to universities are on a royalty-free basis, and those to industry are royalty bearing. It also actively fosters international alliances and collaborative R&D for therapeutic benefit.

In this way, rather than relying on stretched technology transfer offices that may lack the expertise, private IP collectives could compete to attract portfolio management, consolidate portfolio holdings under one roof and improve licensing transparency. Another clear solution to negotiating a fragmented landscape is to 'bundle' related rights to platform technologies through cross-licensing schemes. These cross-licensed rights are then available for

license as a 'package' by a single provider. The management and expertise required for such patent bundling may still best be handled by a company or clearing house, however.

The reasons for patenting an array of methods for the same general technology, such as upstream iPSC production, are multifactorial and do not always indicate a thicket. Inventive steps sometimes address different processes, stages or mechanisms of action that achieve the same function. Alternatively, many patents may seem to focus on the same specific technology in part because working with multiple cell types across different species presents unique challenges. The validity of independent claims to the same iPSC technologies is nonetheless becoming increasingly murky. A splintered approach to developing platform technologies is currently in progress and reflected by the rapidly increasing entanglement of IP among a web of assignees. Although the conditions for a patent thicket exist within the iPSC innovation ecosystem, there are a number of reasons why escalating numbers of dispersed patents may not signify the erection of an anti-commons obstacle.

### Next-generation technology

The first methods of generating iPSCs involved retroviral delivery of reprogramming factors through random integration and were therefore deemed unsafe for therapeutic application. Recent patents claim substantially different methods of and materials for delivering reprogramming factors. For example, small-molecule reprogramming, as recently described in mouse cells<sup>16</sup>, represents a breakthrough technology that may be unaffected by any patent thicket around reprogramming technologies<sup>17</sup>. Unless held to be an obvious progression from previous knowledge, the technology described in these later patents should be able to operate freely with respect to the same technology—production of iPSCs from somatic cells.

**Table 3** Assignees with the most fundamental iPSC patents by forward-citation number

Assignee	Number of citations for most cited patent	Number of patents with >5 forward citations
Kyoto University	104	9
University of California	9	2
General Hospital Corp.	20	2
Harvard College	20	3
Whitehead Institute of Biomedical Research	18	5
Wisconsin Alumni Research Foundation	37	4
National Institute of Advanced Industrial Science and Technology	3	0
Cellular Dynamics International	12	3
Guangzhou Institutes of Biomedicine and Health	2	0
Advanced Cell Technology	15	5

How the field will develop, and whether next-generation technologies must license extensive portfolios or will avoid infringement of prior patents, remains to be seen. As discussed, the patent diaspora may merge naturally through the cross-licensing of readily available commercially bundled technologies. As therapies or products near market, industry-led mergers and acquisitions or strategic partnerships may also develop a coherent IP infrastructure. Consolidation of the IP landscape, through market trends and industry growth, for better-resourced innovation has yet to take shape. The possibilities suggest that concerns are premature and the iPSC landscape may merely reflect a normal early growth cycle within a dynamic industry.

However, practices can be implemented to foster the incentives the patent system offers while protecting translation from patent mismanagement. Improved transparency through the creation of a centralized IP registry that would track current nonconfidential licensing practices would better inform any freedom-to-operate concerns<sup>2</sup>.

### The wider context of iPSC technologies

The validity and quality of individual claims asserted by patents is not assessed here. Even the most fundamental iPSC patents granted, first in 2008 by Yamanaka in Japan and then in 2010 to Kazuhiro Sakurada in the United Kingdom, have not escaped challenges to validity by patents held by Rudolf Jaenisch<sup>18,19</sup>, which significantly predate them. Nonetheless, the patent landscape cannot be understood in isolation from its greater legal, scientific and regulatory context, in which unique challenges exist for the cell therapy industry as a whole. An appreciation of these issues is vital to the understanding of the complexity of claiming proprietary rights in living biomaterials from donors.

A number of recent judicial decisions in the United States and Europe potentially limit the scope of IP rights claimed for stem cells, but iPSC-related inventions remain actively patent eligible. Notably, however, the EU awaits express judicial confirmation this year that iPSCs do not fall within its broad interpretation of 'human embryo' by inducing germ and other cells into an embryonic-like state<sup>20</sup>.

The US Supreme Court decision in *Association for Molecular Pathology et al. v. Myriad Genetics, Inc., et al.* invalidated patents on isolated DNA sequences, causing concern that patents claiming ownership of a donor's isolated stem cells may be refused on similar

'product of nature' grounds. A challenge to the broad ESC patents held by the Wisconsin Alumni Research Foundation was then reinvigorated in the United States<sup>21</sup>. iPSCs are less susceptible to these patentability concerns. Although more recent methods do not rely on genetic manipulation, complex steps chemically manipulate the cells so that they function in a manner that does not occur in nature. Thus, more akin to recombinant DNA, iPSCs are likely to be patentable as both process and composition-of-matter (cell) claims; unless the latter is deemed legally excluded subject matter. Likewise, because iPSCs are not sourced from human embryos, they present fewer ethical challenges. The controversy surrounding human ESCs has confused and cast aspersions on iPSC patentability. Greater legal clarity, beyond Europe<sup>18</sup>, is required to overcome misperceptions.

### IP barriers and solutions

IP support currently available is likely to be insufficient for the large number of small biotech companies, for which the main issue may be initial stacked licensing fees. Typically, these smaller companies with undeveloped IP portfolios will not have the resources for big pharma-level corporate licensing schemes or litigation challenges from biotech goliaths. It is, however, anticipated that these issues are temporary until more iPSC technologies are ready for commercialization. As a field matures, industry partnerships and buyouts typically consolidate IP and provide necessary resources.

Most collaborative efforts have addressed 'open science' sharing of research materials and data but have neglected to address a need for better transparency in licensing practices. Although it is unlikely that patent holders will altruistically donate patents to a true open-source licensing collective, other commercially reasonable alternatives exist to foster the iPSC business community. Transparent commercial licensing practices could be established through the creation of online regional registries or databases, a global IP exchange or, more comprehensively, informal 'clearinghouses' for holders of stem cell-related patent to register ownership and its nonconfidential trading activity<sup>2,3</sup>. The clearinghouse model has been implemented in the public sector for agricultural biotechnology<sup>2</sup>. New licensing models such as the 'Easy Access IP' initiative (<http://www.easyaccessip.org.uk/>) may be suitable for smaller biotech collaboration with universities<sup>22,23</sup>. Other innovative life science platforms also exist for easy adaption

to create an IP registry and exchange market (for example, <http://www.bioinnovit.com/>).

As the wider IP life-sciences landscape shifts to accommodate genomics and regenerative medicine, other forms of IP are surfacing as alternative sources of commercial value beyond patents. Trade secrets and access to corporate biobanks may become the IP concerns for the future of iPSC innovation and public health.

The lengthy regulatory path to market for cell therapies may cause patents to expire before they can be commercially exploited. Trade secrets prevent the sharing of knowledge but circumvent patent expiry and uncertainty issues. Consequently, trade secrets on improved techniques and technical know-how may become the more viable alternative and are appearing as an IP instrument in current industry practice. Cellular Dynamics International, for example, is a leading assignee of iPSC technology and supplies iPSC lines. The company has adopted trade secrets as a vital part of its IP portfolio<sup>24</sup>.

### Conclusions

Unlike human embryonic stem cells, the pivotal impending legal concern for iPSC technology is not patentability. Instead, it is the potential formation of a patent thicket and mismanaged licensing practices in a field that has begun to privatize at an early preclinical phase. The likelihood of these potential challenges occurring is uncertain. Innovative IP licensing solutions, industry growth cycles and the promise of next-generation breakthroughs suggest that such issues should resolve themselves; however, this *laissez-faire* approach may come at too high a risk for the industry and for global health.

Undeniably, the requisite conditions for patent-thicket formation exist. The complex landscape lends credibility to concerns that it is becoming increasingly difficult to navigate and negotiate access to core technologies from dispersed patent holders. This is exacerbated with uncertain validity of overlapping claims by often still-pending patents. As the technology matures, this should be less of an issue.

A pervasive multisector and multinational collaborative culture has already begun within the iPSC community<sup>25</sup>. The damaging IP barriers that potentially loom in the near future could be avoided if stakeholders could equally embrace collaboration (beyond biobanks and data sharing) around IP. A collaborative approach to innovation could be achieved through well-managed IP-exploitation companies or through informal pooling of interrelated patents for certain core technologies through a



clearinghouse. These actions would go far to remove the perceived threat of IP barriers and better serve public health.

Encouraging low licensing fees cultivates mutually beneficial trading partners. Government and not-for-profit institutions are well positioned to lead the way for iPSC technologies in discouraging obstructive practices. This should include terms and conditions, such as for funding or material transfer agreements, which prohibit or create disincentives for exclusive licensing, trade secrets and unreasonable access fees for biobanks. The implications of current licensing practices are only now beginning to surface. Negotiating cumulative licensing schemes from splintered ownership will become increasingly difficult without collaborative efforts to consolidate and ensure an affordable, transparent approach to innovation.

#### ACKNOWLEDGMENTS

The authors thank B. Larner, F. Wattler and J. Evans (Thomson Reuters) for their invaluable assistance; A. Carr (University of Oxford) and J. Karp (Harvard University and Brigham and Women's Hospital) for their invaluable support and insights; and R. Barker (the Centre for the Advancement of Sustainable Medical Innovation (CASMI)) and M. Morys (CASMI) for their support of the CASMI Translational Stem Cell Consortium. The authors wish to express sincere thanks to the following

organizations that have contributed to the consortium as funding and events partners: Centre for the Commercialization of Regenerative Medicine; Celgene Cellular Therapeutics; Cell Therapy Catapult; CIRM; Eisai; GE Healthcare; Lonza; SENS Research Foundation; TAP Biosystems (now Sartorius Stedim); MEDIPOST Co., Ltd.; MEDIPOST America Inc.; the New York Stem Cell Foundation and the US National Institutes of Health Center for Regenerative Medicine. Additionally, CASMI is a past recipient of funding from the UK Technology Strategy Board to support an investigation into cell therapy regulation. The content outlined herein represents the individual opinions of the authors and may not necessarily represent the viewpoints of their employers. D.B. is subject to the CFA Institute's codes, standards and guidelines and as such, must stress that this piece is provided for academic interest only and must not be construed in any way as an investment recommendation.

#### COMPETING FINANCIAL INTERESTS

The authors declare competing financial interests: details are available at <http://www.nature.com/doi/funder/10.1038/nbt.2975>.

- Bubela, T., FitzGerald G.A. & Gold, E.R. *Sci. Transl. Med.* **4**, 133cm3 (2012).
- Bergman, K. & Graff, G. *Nat. Biotechnol.* **25**, 419–424 (2007).
- Winickoff, D.E., Krishanu, S. & Graff, G.D. *Yale J. Health Policy Law Ethics* **9**, 52–157 (2009).
- Mathews, D.J.H. *et al. Science* **331**, 725–727 (2011).
- Schwartz, S.D. *et al. Lancet* **379**, 713–720 (2012).
- Takahashi, K. *et al. Cell* **131**, 861–872 (2007).
- Yu, J. *et al. Science* **318**, 1917–1920 (2007).
- Mathews, D.J.H., Cook-Deegan, R. & Bubela, T. *Cell Stem Cell* **12**, 508–512 (2013).
- The Hinxton Group. Consensus statement on policies and practices governing data and materials sharing and intellectual property in stem cell science. (The Hinxton Group, 2010).
- Rao, M.S. *Stem Cell Res. Ther.* **4**, 98 (2013).
- Heller, M.A. & Eisenberg, R.S. *Science* **280**, 698–701 (1998).
- BioInformant Worldwide. Complete 2012–13 induced pluripotent stem cell industry report (BioInformant Worldwide, LLC, 2014).
- Cyranoski, D. *Nat. Med.* **19**, 510 (2013).
- United States of America v. Regenerative Sciences, A Corporation, LLC et al.* No. 12–5254, United States Court of Appeals for the District of Columbia Circuit, 4 February 2014.
- UK Intellectual Property Office Patent Informatics Team. Stem cells: the UKSCN patent watch landscape (Intellectual Property Office, Newport, UK, 2012).
- Hou, P. *et al. Science* **341**, 651–654 (2013).
- Feldman, R. & Furth, D. *Stanford J. Law Sci. Pol.* **3**, 17–34 (2010).
- Simon, B.M., Murdoch, C.E. & Scott, C.T. *Nat. Biotechnol.* **28**, 557–559 (2010).
- Japan Patent No. 2008–131577, 2008; UK Patent No. GB2450603, 2010; US Patent No. 7,682,828, 2010; US Patent Application Patent No. 12/703,015, 2010; US Patent Application Patent No. 12/703,015, 2010.
- International Stem Cell Corporation v. Comptroller General of Patents*, Patents Court, London, UK, Case No. [2013] EWHC 807 (Ch.); Referral Case C-364/13 (Court of Justice of the European Union, 17 April 2013).
- Consumer Watchdog v. Wisconsin Alumni Research Foundation* (Fed. Cir. 2013).
- House of Commons Science and Technology Committee. Bridging the valley of death: improving the commercialisation of research. Eighth report of session 2012–13 (Report no. HC 348) (House of Commons Stationery Office Limited, London, 2013).
- Moran, N. *Nat. Biotechnol.* **31**, 376 (2013).
- Initial public offering, filed with the Securities and Exchange Commission, Washington DC, on 3 June 2013, Registration no. 333.
- Webb, S. *Nat. Biotechnol.* **27**, 977–979 (2009).

# PERSPECTIVES

## OUTLOOK

### Can open-source R&D reinvigorate drug research?

Bernard Munos

**Abstract** | The low number of novel therapeutics approved by the US FDA in recent years continues to cause great concern about productivity and declining innovation. Can open-source drug research and development, using principles pioneered by the highly successful open-source software movement, help revive the industry?

Open-source research, which started as a counterculture movement in the software industry 15 years ago, has since grown into a business model whose best-known product, Linux, has become a credible alternative to Microsoft's Windows. Now, with biology increasingly becoming an information-orientated science, some have suggested that what worked for software might be part of the answer to the spiralling cost of drug R&D. With this in mind, this article examines the relevance to pharmaceutical R&D of the open-source model developed by the software industry. In this context, open-source no longer refers to source code, but instead to the open origin of contributors.

Open-source R&D has already made inroads into bioinformatics and research tools for drug hunters. However, key differences between software and biology, such as regulatory requirements, have limited its application to drug development. Nevertheless, in the past 5 years a new breed of organizations called public-private partnerships (PPPs) have adapted the open-source concept and combined it with outsourcing to create a new, low-cost business model, which they have applied with encouraging results to the discovery of new treatments for neglected diseases.

Advances in data mining, visualization and networking now make it feasible to go one step further. It is possible to offer scientists a computerized toolbox that lets them harness the creativity of numerous volunteers to address the key questions that are holding back innovation. For example, what is the aetiology of a disease? What are the pathways

involved? What are the better targets? Once these questions are answered, laboratory and clinical studies can be outsourced to institutions with the requisite capacity through the help of matchmaking software.

The resulting model is a hybrid in which a part of R&D is open-sourced while the rest is outsourced. To function, however, it needs strong project leadership and expertise in the minutia of drug R&D, which mostly exist in big pharmaceutical firms. This suggests that, far from being a threat to conventional drug R&D, open-source could be a way to leverage big pharma's capabilities in order to tackle challenges that the blockbuster model cannot address economically, such as neglected diseases. As pharmacogenomics takes hold, it might also be a way to address market niches that cannot support blockbusters.

#### A brief primer on open-source

Open-source R&D is a novel approach to research that lets scientists join hands freely across organizations, disciplines and borders to solve problems in which they share an interest. The movement's icon is Linux, the operating system started in the early 1990s by student Linus Torvalds, who used the nascent Internet to circulate it to fellow computer enthusiasts. Soon they were busy adding features and improving the code, with Torvalds overseeing the process. Fifteen years later, this grassroots experiment has blossomed into a new culture that is spreading to other disciplines. It is most prominent in computer software development, for which dedicated websites such as [SourceForge](#) or

[Subversion](#) help over a million people collaborate on more than 100,000 projects. But other areas, such as life sciences, have spawned open-source initiatives of their own.

The impetus to create open-source software often comes from developers looking for challenge. They agree on an attractive project, form a team and produce a 'bare-bones' program with basic functionality. Then, they offer it at no cost on the Internet under a public-domain license (there are many different types of open-source license; some, notably the 'copyleft' or General Public License (GPL), require those who download a program to share any improvements they make). If the project draws interest, others add features and post their code on the project's webpage for fellow programmers to critique. New code of sufficient quality is added to the authorized version of the program.

Open-source's chief benefit is to cross-fertilize minds and tap creativity quickly, cheaply and on a scale that is beyond the reach of scientists working in the 'ivory towers' of academia or behind the 'corporate moats' of industry. Hollingsworth<sup>1,2</sup> has shown that innovation spikes when diverse minds interact frequently in an unstructured manner. By drawing talent from all around the world, open-source research takes these dynamics to a new scale. And by making innovation immediately available to all, it speeds up the accumulation and application of knowledge.

Outsiders are often puzzled by the open-source idea. Why would anyone work for free? Simply put, because some people value non-cash compensation more than money. They volunteer their expertise to satisfy idealism or curiosity, seek new challenges, hone skills, build a reputation or enhance careers. Feldman<sup>3</sup> quotes the example of Australian programmers who, within hours of Netscape's release of its browser code, attached an 'add-on' to enable secure internet transactions. No money changed hands, but the authors received respect from the programming community and the satisfaction of turning out an elegant and useful piece of software.

Companies are learning to use open-source to their advantage, and many now allow their employees to participate on company time. They might use it to gain market share against entrenched competitors, or

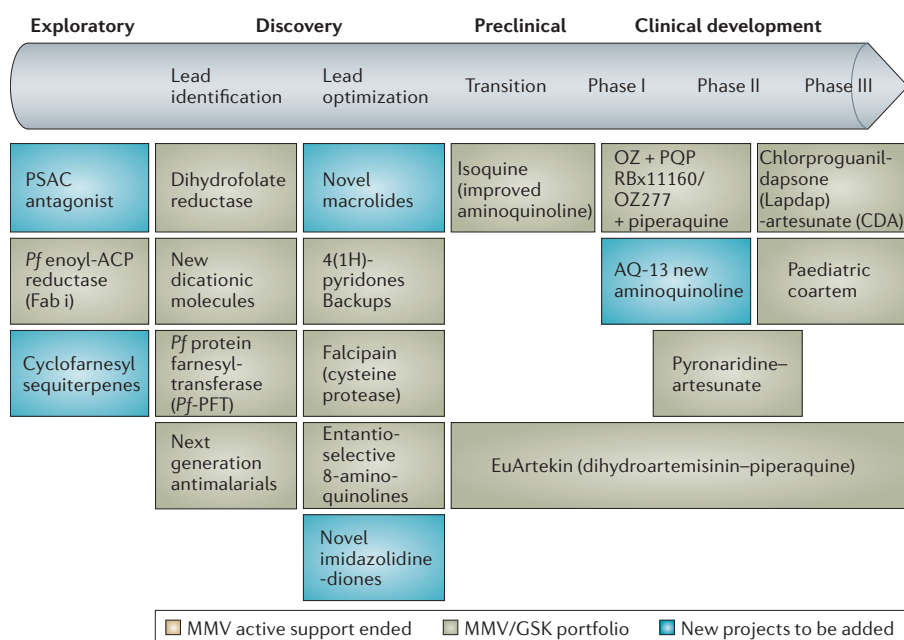


Figure 1 | **Portfolio of the Medicines for Malaria Venture.** PSAC, plasmodial surface anion channel.

to entice developers to create applications for their product, possibly in the hope of turning it into a 'platform'. Some of them have been quite successful at turning open-source into profits. Red Hat, for instance, has attained a US\$5-billion market cap from selling support services for Linux.

### Can it work for drugs?

If biomedical scientists could adapt the open-source model, it could make a huge difference to such projects as developing drugs for neglected diseases, for which needs are great but funds are scarce<sup>4</sup>. Only 10% of R&D resources are spent on illnesses that represent 90% of the burden of disease. Open-source drug R&D might not change that equation, but could make it possible to get much more from that 10%.

There are, however, significant barriers to the deployment of open-source approaches to drug R&D<sup>5</sup>. One is economic. All it takes to write open-source software is a laptop and an internet connection. With drug research, someone must pay for laboratory expenses and clinical trials. And the costs are high, at more than US\$800 million for the discovery and development of a novel drug by most estimates.

Research dynamics between the two industries also differ. Software development does not have a discovery phase. Once the objective is set, programmers set to work and make steady progress towards their goal. By contrast, drug discovery cannot flourish until a certain amount of knowledge about the target disease

has been accumulated. That knowledge acquisition can take years or decades, with no way to know at the outset whether the store of knowledge at hand is nearly sufficient or will require years of painstaking additional research before innovation can thrive.

Software development is also simpler: it spans only a few disciplines and has no equivalent to clinical trials. For the most part, a single programmer can master all the skills needed to write a program from start to finish. By contrast, drug development requires coordination of multiple specialties with little overlap. Biomedical knowledge, which grows at the rate of 1,000 publications per day, must be peer-reviewed and replicated before it is accepted. All this is slow and enormously expensive.

Drug R&D can go off-track more easily than software programming. Biologists can get mired in the complexity of biology without ever making much progress towards a drug — chemists handed the wrong target cannot do much good no matter how hard they try; inadequate toxicology can derail a compound late in development, or even after launch. One misstep along the way can render all downstream work useless.

In contrast to drug developers, software publishers are lightly regulated. They do not need FDA approval. The quality standards they face are far less onerous than the minutia of Good Laboratory Practice (GLP), Good Clinical Practice (GCP) and Good Manufacturing Practice (GMP). One sloppy programmer seldom jeopardizes

the achievements of others, and errors can be patched without requiring the rewrite of the whole program. With drugs, one careless worker can compromise years of work costing tens of million of dollars.

Finally, the two industries follow different intellectual property regimes. Software is protected by copyrights that arise automatically as code is written, even if nothing is filed. Drug research is protected by patents that are costly to file and maintain, and for which meeting the legal standards that define innovation is much harder.

### Open-source biomedical research

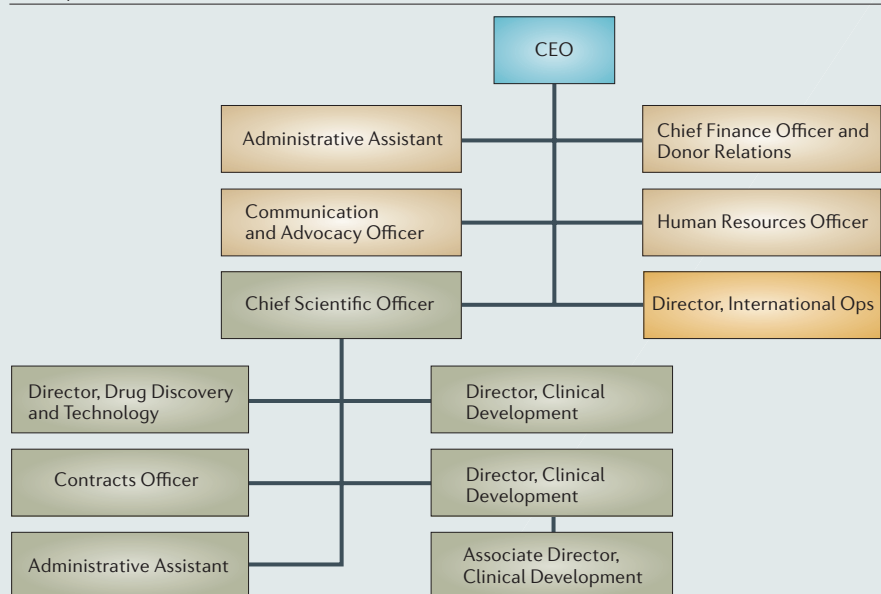
**Early efforts.** Despite these differences, the open-source idea has entered biomedical research<sup>6</sup>. The first inroads were made in bioinformatics<sup>7,8</sup>, as might have been expected. These efforts resulted in a collection of programs such as **Biojava**, **BioPerl**, **BioPython**, **Bio-SPICE**, **BioRuby** and **Simple Molecular Mechanics for Proteins**<sup>9</sup>, and inspired other initiatives such as the Human Genome Project, the SNP Consortium, the Alliance for Cellular Signaling, BioForge, GMOD and Massachusetts Institute of Technology's BioBricks (some of these have the transparency and feel of open-source, although the resources needed to get involved do not allow all volunteers to participate; however, we still call them 'open-source').

**An old idea.** One could argue that there has long been an active, if invisible, collaborative process akin to open-source in drug development, as, for some diseases, half of all prescriptions are for off-label uses<sup>10</sup>. Somehow, physicians share their ideas and experiences informally to uncover novel uses for existing medicines. For instance, oncologists routinely use drugs approved for one kind of cancer to treat other types. In a recent study, DeMonaco<sup>11</sup> found that 59% of drug therapy innovations were discovered by practicing clinicians via field discovery. The way by which physicians uncover these new indications is quick and inexpensive compared with Phase III trials. From an economic and medical standpoint, there would be merit in exploiting these clinical observations and sharing them with physicians as a complement to, or replacement for, some of the traditional clinical development.

**Public-private partnerships.** Taking a different approach, a new kind of organization, known as the public-private partnership (PPP), has recently developed a clever virtual business model that emulates the collaborative features of the open-source concept<sup>12</sup>. An example is



## Box 1 | The Medicines for Malaria Venture

**Management team**

The Medicines for Malaria Venture (MMV) is run by a staff of 13. Its CEO reports to a Board of 12 Directors who represent funding organizations. An Expert Scientific Advisory Committee, which includes chemists, biologists, clinicians, malariologists and drug development experts, advises on project selection and research strategy. The Management Team's responsibilities are to:

- Encourage the submission of research proposals
- Select proposals and negotiate with partners
- Set up project-management teams and monitor progress
- Organize manufacturing and marketing
- Earn appropriate returns from marketed products
- Raise funds
- Communicate with government agencies

**Backers**

- Gates Foundation
- BHP Billiton
- ExxonMobil
- Global Forum for Health Research
- International Federation of Pharmaceutical Manufacturers Associations
- Netherlands Minister for Development Cooperation
- Rockefeller Foundation
- Swiss Agency for Development and Cooperation
- United Kingdom Department for International Development
- United States Agency for International Development
- World Bank
- Wellcome Trust
- World Health Organization

the **Medicines for Malaria Venture** (MMV), which was established in 1999 to discover and develop new, affordable antimalarial drugs. Established as a nonprofit entity with a staff of only 13 people, it has assembled a portfolio of 19 projects ranging from discovery to Phase III (FIG. 1).

MMV gets its projects through open calls — anyone with an idea can contribute. An Expert Scientific Advisory Committee reviews the submissions and selects the projects that will be funded. Each is managed by a project manager who outsources the

R&D to a network of 300 scientists at 40 institutions (universities, big pharma, biotech and research institutes). Funding comes from public and philanthropic partners (BOX 1). After each step, the Scientific Advisory Committee reviews the data and decides whether to proceed or terminate the project. MMV's cumulative spend from 2000 through 2005 is about US\$100 million, 90% of which funded actual research. MMV plans to outsource manufacturing to low-cost partners, sell drugs at cost to developing countries, and market them through partners in developed

markets (for example, to treat travellers). Its alliance with GlaxoSmithKline supports 25 scientists funded equally by the partners.

The **Initiative on Public-Private Partnerships for Health** reckons that there are about 24 PPPs engaged in drug and vaccine R&D (TABLE 1). Most of them were created in the past 7 years and share a common profile<sup>13</sup>. First, they focus on neglected diseases. Second, they operate as virtual drug companies, with a small staff getting project ideas from outside, vetting them through a committee of experts and outsourcing R&D to a network of institutions. Third, they manage growing portfolios of projects ranging from discovery through to Phase III trials. Fourth, they have been able to function on lean budgets with a cumulative spending that seldom exceeds US\$50 million. This makes them attractive vehicles to fund research in areas that are not economical for traditional drug R&D.

By the end of 2005, PPPs had attracted funding in excess of US\$1.5 billion. Foundations have given about US\$1.15 billion (with the Gates Foundation alone contributing more than US\$950 million), governments US\$244 million and private entities US\$36 million. In addition, donors have committed another US\$3.5 billion which will be disbursed as needed by **The Global Fund to Fight AIDS, Tuberculosis and Malaria**.

**Open-source versus alliance networks.** It can be argued that the 25,000 alliances wrought by the 8,000 pharma and biotech companies over the past 15 years add up to a vast open-innovation system that mimics the collaborative features of the open-source model. Some scholars have countered, however, that alliances are less effective than open-source research at promoting innovation. This is because open-source networks are richer in 'weak links' (loose relationships), whereas alliances pride themselves on the strength of the connections between partners. DeBresson<sup>14</sup> has shown that weak links bring novel ideas into the fray whereas strong links tend to reinforce orthodoxies.

**PPPs and big pharma.** TABLE 2 lists some of the projects and organizations coordinated by PPPs. As can be seen, GlaxoSmithKline features prominently, with Bristol-Myers Squibb, Novartis, Bayer, Sanofi-Aventis and Ranbaxy involved to a lesser extent.

**Lessons learned**

PPPs have advantages and drawbacks compared with traditional R&D<sup>15</sup>. Advantages include the following.

Table 1 | **Public–private partnerships engaged in drug and vaccine development**

Name	Focus	Year created
Aeras Global TB Vaccine Foundation	Tuberculosis	1997
BIO Ventures for Global Health	Biotech drugs for neglected diseases	2004
Consortium for Industrial Collaboration in Contraceptive Research	Development of new contraceptives	1995
Contraceptive Research and Development	Improving reproductive health in developing countries	1986
Dengue Vaccine Project	Dengue fever	1989
Drugs for Neglected Diseases Initiative	Sleeping sickness, visceral leishmaniasis, Chagas disease	2003
European Malaria Vaccine Initiative	Malaria	1998
Gates Foundation–UNC Partnership for Development of New Drugs	African trypanosomiasis, leishmaniasis	2000
Global Alliance for TB	Tuberculosis	2000
Global Microbicide Project	New microbicides for women	2000
Human Hookworm Vaccine Initiative	Hookworm	2000
Infectious Disease Research Institute	Tuberculosis, leishmaniasis, Chagas disease, malaria, leprosy and Buruli ulcer	1993
Institute for One World Health	Visceral leishmaniasis, cutaneous leishmaniasis, Chagas disease, paediatric secretory diarrhoea	2000
International AIDS Vaccine Initiative	AIDS	1996
International Partnership for Microbicides	HIV	2002
Japanese Pharmaceutical, Ministry of Health, WHO Malaria Drug Partnership	Malaria	1999
Lapdap Antimalarial Product Development	Malaria	1998
Lassa Fever Initiative	Lassa fever	2001
Malaria Vaccine Initiative	Malaria	1999
Medicines for Malaria Venture	Malaria	1999
Meningitis	Meningitis	2001
Microbicides Development Programme	HIV	2001
Pediatric Dengue Vaccine Initiative	Dengue	2001
PneumoADIP	<i>Pneumococcal</i> vaccines	2004

UNC, University of North Carolina; WHO, World Health Organization.

**Agility.** Virtual R&D makes it easier to terminate projects that no longer look promising. The project manager does not have to deal with entrenched advocates manoeuvring to save their project or move it underground.

**Creativity.** PPPs enable experts from different countries, specialties and styles of thought to leverage each other's ideas. They harness the problem-solving skills of a much greater population than is typically available to traditional research organizations.

**Focus.** PPPs focus on one or few diseases. This helps them build deep expertise for better decisions (for example, target selection).

**Risk sharing.** The open-innovation model of PPPs makes it easier for scientists to collaborate on pre-commercial research such as biomarkers or cell signalling.

**Affordability.** PPPs lower the critical mass required to be a pharmaceutical company. By leveraging external expertise and capabilities, they allow small organizations to do much of what was once the domain of large companies.

**Impact.** PPPs engage scientists in developing nations who have first-hand experience in many neglected diseases. It helps them build their clinical research capacity, which in turn leverages the effectiveness of their public health systems.

**Speed.** Lean PPPs can decide quickly, partly because they do not have layers of committees to satisfy. In addition, because they tap their partner's unused capacity, they can advance swiftly as there is often a qualified laboratory somewhere that can do the work without having to wait in someone else's queue.

There are also some disadvantages to PPPs, which include the following.

**Funding.** US\$5 billion has been committed to PPPs (\$1.5 billion disbursed). However, despite the thriftiness of PPPs, there is concern that these funds will be stretched as more projects move into late, expensive clinical development.

**Sustainability.** PPPs have not demonstrated the sustainability of their business model. Some of their projects come from companies that had shelved them because of insufficient commercial prospects. To survive, PPPs will need to replenish their portfolios. There are also worries that, in some areas of science, the pool of contributors might be too thin to perform the work that must be done.

TABLE 3 shows that the PPP R&D model has worked reasonably well. Some of this success comes from targeting low-hanging fruits in diseases that have long been neglected, but it also suggests that the PPP model can be a potent tool in finding new cures. Whether the PPP business model becomes a transformational force or remains a non-threatening niche depends on how it ultimately performs against traditional pharmaceutical R&D. To succeed, it must go beyond tools and software and tackle large projects where it will rival the big firms that are helping it today. Yet, this rivalry need not be a zero-sum game. On the contrary, there is a place for collaborative and proprietary research in drug R&D, just as in software<sup>16</sup>. If open-source drug R&D takes hold, what will probably emerge is not the replacement of one model by another, but an ecology in which big pharma, biotech and collaborative research compete and collaborate at the same time, feeding off each other synergistically, while moving towards therapies along their own distinctive paths.

#### **A template for open-source drug R&D**

Can the PPP model succeed beyond neglected diseases? To answer this, it helps to break down drug R&D into knowledge-based activities and rule-based tasks.

Knowledge-based work requires lots of intelligence and intuition, but little infrastructure. Examples include identifying targets, understanding metabolic networks, and designing clinical trials or computerized disease models. It is about scientists leveraging each other's ideas, and using tools to gain deeper insights that might lead to breakthroughs. This work is ideally suited to the open-source model.

Rule-based work requires physical assets (laboratories, equipment, patients and so on) and money. It is tightly scripted and must conform to rigid regulatory requirements. It is about organization, discipline and implementation. Examples include toxicology studies, Chemistry Manufacturing and Controls (CMC) studies, and the conduct of clinical trials. Rule-based work is ideally suited to outsourcing, and much of it is already outsourced to contract research organizations.

This division of labour suggests a business model template in which part of the R&D value chain is open-sourced, while the rest is outsourced, with the following features.

#### Template features: operating principles

**Open-sourcing.** The open-source part of our model should allow anyone who can contribute to join. Volunteers should be able to log on to a website, find the page(s) that matches their area of expertise, peruse challenges to be solved, review others' contributions, download computerized tools and start working towards contributions of their own. As they progress, they can publish their findings in scientific journals and discuss their insights in on-line forums. Over time, the better ones will gain authority and become the *de facto* leaders of their open-source community.

**Outsourcing.** Work to be outsourced should be posted on a website for all to see. Scientists and organizations qualified for the job can bid, and the sponsor picks the best candidate for each task.

#### Template features: procedures

**Governance.** Three decision-making bodies provide leadership and guidance: the Board of Directors, the Steering Committee and the Scientific Advisory Committee. The Board of Directors includes senior executives and outsiders who represent shareholders and stakeholders. It approves strategy and ensures that management performance is consistent with the organization's mission. The Steering Committee

Table 2 | **Public-private partnerships and their partners**

Project	Industry partner	University/public health partner	Other PPP
<b>Medicines for Malaria Venture</b>			
Improved 4-aminoquinoline	GlaxoSmithKline	University of Liverpool	
Farnesyl transferase inhibitors	Bristol-Myers Squibb	University of Washington	
Manzamine derivatives		University of Mississippi	
Cysteine protease inhibitors	GlaxoSmithKline	UCSF	
Fatty acid biosynthesis inhibition		Texas A&M, A. Einstein, Jacobus	
Pyridone	GlaxoSmithKline		
New di-cationic molecules	Immtech	University of North Carolina	
Dihydrofolate reductase inhibition	Biotec Thailand		
Artesunate derivatives	GlaxoSmithKline, Shin Poong		TDR, DNDi
Artemisone	Bayer	University of Hong Kong	
Synthetic peroxide	Ranbaxy	University of Nebraska	
Intravenous artesunate		Walter Reed	
Coartem in infants	Novartis		TDR
<b>TB Alliance</b>			
Pyridones and quinolizines	Taejon, Yonsei		
Isoniazid analogue		Wellesley College	
PA-824		NIH, Johns Hopkins	
Mocifloxacin	Bayer	CDC, Johns Hopkins	
<b>Institute for One World Health</b>			
Paromomycin			TDR
Azole		Yale	
<b>TDR</b>			
Miltefosine	Zentaris		
Oral efloornithine	Aventis		

CDC, Centers for Disease Control and Prevention; DNDi, Drugs for Neglected Diseases Initiative; NIH, National Institutes of Health; TDR, UNICEF-UNDP-World Bank-WHO Special Programme for Research and Training in Tropical Diseases.

is a group of senior executives that rules on important operational issues such as fundraising, budgets, project funding, key hires and selection of partners.

It also approves recommendations from the Scientific Advisory Committee. The Scientific Advisory Committee (SAC) is a group of external experts from academia and industry. It sets R&D strategy, proposes new projects, reviews existing ones and recommends termination of those that no longer deserve support.

**Scope.** This template calls for focusing on single diseases or related illnesses. An organization working on unrelated diseases should establish separate websites for each one.

**Projects origination.** There is a permanent open call for new projects. Scientists are invited to submit ideas online for review by the SAC.

**Portfolio management.** The SAC is responsible for maintaining an adequate and balanced pipeline of projects.

**Project management.** Each project is managed by a Project Team led by a member of the organization, and staffed by external experts in drug discovery, clinical research and regulation. The Project Team is responsible for developing the budget and timeline, overseeing outsourced tasks and ensuring compliance with GxP. One of its crucial duties is selecting what will be open-sourced



Table 3 | **Public–private partnerships lower the critical mass required to discover and develop new cures**

Organization	Focus	Staff	Pipeline				
			Number of projects	Discovery	PK	Clinical	Cumulative spending through 2005 (US\$ million)
Medicines for Malaria Venture	Malaria	13	19	12	1	6	103
TB Alliance	Tuberculosis	18	12	8	1	3	20
Drugs for Neglected Disease Initiative	Trypanosomiasis, visceral leishmaniasis, Chagas disease	36	20	9	4	7	20
OneWorld Health	Leishmaniasis, malaria, Chagas disease, diarrhoeal diseases	40	5	1	3	1	?
International AIDS Vaccine Initiative	HIV/AIDS	169	6	–	1	5	120
Malaria Vaccines Initiative	Malaria	32	10	4	2	4	?

Source: Annual reports. PK, pharmacokinetics.

and what will be outsourced. The project leader is accountable for generating the data used to decide whether to fund the next step. Commitment to a project is limited to the current step, until the data warrants committing funds for the next one. Open-sourced tasks are posted on the project's website, each on its own page, and outsourced ones are posted on a companion matchmaking website such as **Innocentive**, or **Sciencetour**. Outsourcing bids are reviewed by the Project Team, which issues recommendations to the SAC.

**Intellectual property ownership.** There is often a misperception that open-source initiatives are hostile to patents and bent on putting discoveries in the public domain. The reality is more nuanced. Most open-source activities occur at a pre-commercial R&D stage, when the ideas and hypotheses debated fall short of the legal standards that define inventions in patent law. They are an on-going scientific conversation that can be likened to a global instant-messaging system linking scientists interested in a topic. In that sense, open-source is no more threatening to patents than other forms of scientific publishing. A scientist who engages in that conversation and comes up with an idea that

can lead to a patentable invention will need to exercise caution with disclosures until the invention has been reduced to practice and patent applications have been filed, just as would be necessary in a traditional research setting. It is generally accepted that open communication promotes advancement of science, but needs to be balanced by the need to protect the rights of inventors. The same applies to open-source activities.

#### Template features: tools

**The discovery toolbox.** As of February 2006, 349 genomes have been published and another 1,575 are being sequenced. A new generation of smart, computerized tools is becoming available to mine data, comb the literature, map metabolic networks, perform *in silico* modelling, visualize binding sites, identify chemical leads, design molecules and predict toxicity. These tools should be packaged into a convenient toolbox, together with access to major databases, and offered to volunteers willing to contribute their expertise.

**Outsourcing software.** Several programs already exist to match projects with talent and capacity. Two examples are Sciencetour, a free e-marketplace that allows companies

to post tasks, and experts to register their skills, and Innocentive, an online problem-solving tool that lets a company post a challenge with a reward: whoever finds the solution gets the money.

#### Template features: costs

PPPs have been able to function on very low budgets for several reasons (TABLE 4). First, they have few people, low overhead costs and no fixed assets. They rely on someone else's unused capacity, and the market seems to price such capacity at marginal instead of full cost. Second, they outsource much of their work where it is cheaper to do so and do most of their trials in developing countries. Third, they concentrate on infectious diseases for which costs are lower. Fourth, they receive in-kind donations.

#### Will it work?

Despite the promise of open-source drug R&D, both its pioneers, and the veterans of open-source software, point to several potentially troublesome issues that could affect the success of the open-source model.

**Availability of talent.** Typical open-source projects do not require a large number of contributors. Data from the software industry suggests that the ideal number ranges from 6 to 20 people. Yet much of the drug R&D expertise resides in an industry that has a strong proprietary culture. Employees are routinely asked to assign their intellectual output, including that created on their own time, to their employers<sup>17</sup>. This could stifle talent supply in key areas. Two developments, however, might give open-source drug R&D the permanent talent pool it needs. First, thousands of highly trained pharmaceutical scientists are nearing retirement and might

Table 4 | **R&D costs for public–private partnerships (US\$ million)**

Stage	MMV	TB Alliance	DNDi	IAVI	Big Pharma
Discovery and PK	8.3	18.6	16.2	20.0	26.0
Phase I	1.6	0.6	Unpublished	2.0	15.2
Phase II	1.2	3.4	Unpublished	5.0	23.5
Phase III	9.5	22.6	Unpublished	30.0	86.3
Total clinical	12.2	26.6	24.2	37.0	125.0

Source: REF. 19. DNDi, Drugs for Neglected Diseases Initiative; IAVI, International AIDS Vaccine Initiative; MMV, Medicines for Malaria Venture.

welcome the opportunity to put their skills to good use. Second, drug companies might be persuaded to ease restrictions on their employee's involvement. There is indeed little conflict of interest in being a cancer scientist by day and an anthrax researcher at night, and firms might gain valuable goodwill from letting employees seek cures for diseases in which they have no interest.

**Availability of data and standards.** Open-source scientists cannot accomplish much unless they can access data. Biological data is plentiful and getting richer, with terabytes of genomic and metabolic data continuously being added to the pool. Chemical and structural data, on the other hand, are more scarce. In addition, the formats used to handle these data are still evolving. Biologists use a reasonably small number of them, but chemists are further from such consensus. Both the lack of standards and the scarcity of data in certain areas can cause problematic choke points in an open-source R&D effort.

**Availability of tools.** Open-source scientists need open-source tools to practice their craft. Until recently, such tools were plentiful in bioinformatics, but less so in chemistry, which has long been dominated by commercial software. This is changing. The 2004 launch of **PubChem** has brought online a powerful suite of tools that allows scientists to connect chemical information with biomedical research and clinical information in an unprecedented way. Non-profit scientists can now access small-molecule high-throughput screening, chemistry and informatics on a scale previously available only to industry. They can even get grants to turn their online discoveries into assays for high-throughput screening<sup>18</sup>. Other tools such as **eMolecules**, **Jmol** or the **Chemistry Development Kit** are adding powerful chemical search and visualization capabilities to the open-source scientist's toolbox.

**Intellectual leadership.** Just as putting ingredients into a vat does not necessarily cause them to react, connecting smart people online does not guarantee they will produce anything valuable. In both cases, a catalyst is needed. For open-source drug R&D, the presence of a subgroup of highly innovative contributors who can tune in the on-going conversation and fuel it with their own creative insights acts as such a catalyst. Without it, the conversation could remain shallow and fizzle out.

**Momentum.** Enticing people to join is a challenge. A good website helps, but it's not

enough. As Darren Carroll, former CEO of Innocentive, puts it, "If you build it, they will not come!". It takes a sustained effort to get the word out and build trust with stakeholders. It also takes a leader who can connect with people, understand their motivation and foster trust. Linux attracts thousands of contributors because they identify with Torvalds' ideals and trust him to do the right thing. Open-source drug R&D must build such leaders.

**Web interface.** The design of the project's website is crucial. It must be engaging and appeal to visitors' curiosity. They must be able to quickly find the pages that match their interests, download the toolbox, and be 'up-and-playing' in minutes.

**Quality assurance/quality control.** When something as complex as drug R&D gets parceled out around the world, quality assurance can become an issue. Oversight, due-diligence, audits, good practices and prior experience can be used to ensure quality. International Organization for Standardization standards could also help in the future.

**Selectivity.** Not all projects will be equally suitable. Cancer might draw contributors, but hair loss might not.

### Conclusion: a new ecology of drug R&D?

Is there still room for big pharma in open-source R&D? One must stress that 'virtual' does not mean 'leaderless'. To succeed, open-source R&D will need deep expertise in the minutia of drug R&D, which today resides overwhelmingly in the pharmaceutical industry. There might be many volunteers, but they must be shepherded towards a goal. Such stewardship is a core competency of pharmaceutical companies. Our model is not a substitute for them, but a way to leverage their capabilities to tackle unmet medical needs, such as the diseases of poverty, orphan diseases and niche markets. Pharmaceutical companies stand to gain from co-opting the open-source model and allowing it to flourish in 'coopetition' with traditional R&D, to handle the diseases or R&D steps for which it is best suited.

Bernard Munos is at Eli Lilly & Co., Lilly Corporate Center, 1085, Indianapolis, Indiana 46285, USA.  
e-mail: bhmunos@stanfordalumni.org

doi:10.1038/nrd2131

Published online 18 August 2006

- Hollingsworth, J. R. & Hollingsworth, E. J. in *Practicing Interdisciplinarity* (eds Weingart, P. & Stehr, N.) 215–244 (Univ. Toronto Press, Toronto, 2000).
- Hollingsworth, J. R. in *Creating a Tradition of Biomedical Research* (ed. Stapleton, D.) 17–63 (Rockefeller Univ. Press, New York, 2004).

- Feldman, R. The open-source biotechnology movement: is it patent misuse? *Minn. J. L. Sci. Tech.* **6**, 1 (2004).
- Cukier, K. N. Community property: Open-source proponents plant the seeds of a new patent landscape. *Acumen* **1**, 54–60 (2003).
- Rai, A. Open and collaborative research: a new model for biomedicine. Duke Law School, Legal Studies Research Paper Series, Research Paper **61** (2004).
- Maurer, S. New institutions for doing science: from databases to open biology. Presented at University of Maastricht, November 24–25 (2003).
- DeLano, W. L. The case for open-source software in drug discovery. *Drug Discov. Today* **10**, 213–217 (2005).
- Geldenhuys, W. J., Gaasch, K. E., Watson, M., Allen, D. D. & Van der Schyf, C. J. Optimizing the use of open-source software applications in drug discovery. *Drug Discov. Today* **11**, 127–132 (2006).
- Eisenmenger, F., Hansmann, U. H. E., Hayryan, S. & Hu, C. An enhanced version of SMMP — open-source software package for simulation of proteins. *Computer Phys. Comm.* **174**, 422–429 (2006).
- An open-source shot in the arm. *The Economist* (12 June 2004).
- DeMonaco, H. J., Ali, A. & Von Hippel, E. The major role of clinicians in the discovery of off-label drug therapies. MIT Sloan Working Paper 4552-05 (2005).
- Maurer, S. M., Rai, A. & Sali, A. finding cures for tropical diseases: is open-source the answer? *PLoS Med.* **1**, e56 (2004).
- Gardner, C. & Garner, C. Technology Licensing to nontraditional partners: non-profit health product development organizations for better global health. *Industry Higher Education* **19**, 241–247 (2005).
- DeBresson, C. & Amesse, F. Networks of innovators: a review and introduction to the issue. *Res. Policy* **20**, 363–379 (1991).
- Nwaka, S. & Ridley, R. Virtual drug discovery and development for neglected diseases through public-private partnerships. *Nature Rev. Drug Discov.* **2**, 919–928 (2003).
- Hope, J. *Open-Source Biotechnology*. Ph.D. Thesis, Australian National Univ. (2004).
- Stahl, M. T. Open-source software: not quite endsville. *Drug Discov. Today* **10**, 219–222 (2005).
- Collins, F. S. The NIH Roadmap: new pathways to discovery — empowering small molecule research. *Office of Portfolio Analysis and Strategic Initiatives* (National Institutes of Health, Bethesda, 2006).
- Towse, A. & Renowden, O. in *Combating Diseases Associated with Poverty: Financing Strategies for Product Development and the Potential Role of Public-Private Partnerships* (eds Widdus, R. & White, K.) [online], <[http://www.globalforumhealth.org/filesupld/ippph\\_cd/06.PDF](http://www.globalforumhealth.org/filesupld/ippph_cd/06.PDF)> [Initiative on Public-Private Partnerships for Health, London, 2004].

### Acknowledgements

I thank A. Tashjian (Harvard School of Public Health and Harvard Medical School), B. Smith (Center for Biosecurity, University of Pittsburgh Medical Center) and M. Munos (Gardner Carton & Douglas) for valuable feedback on previous versions of the manuscript.

### Competing interests statement

B.M. works for Eli Lilly & Co., which has sponsored the Scientist and Innocentive ventures mentioned in this article. This declaration of potential competing financial interests is also available in the Web version.

### FURTHER INFORMATION

SourceForge: <http://sourceforge.net/>  
 BioForge: [www.bioforge.net](http://www.bioforge.net)  
 PubChem: <http://pubchem.ncbi.nlm.nih.gov>  
 eMolecule: [www.emolecules.com](http://www.emolecules.com)  
 Jmol: <http://jmol.sourceforge.net>  
 CDK: <http://almost.cubic.uni-koeln.de/cdk>  
 Emboss: <http://emboss.sourceforge.net/>  
 Medicines for Malaria Venture: [www.mmv.org](http://www.mmv.org)  
 The Initiative on Public-Private Partnerships for Health: [www.ippph.org](http://www.ippph.org)  
 Global Fund to Fight AIDS, Tuberculosis and Malaria: [www.theglobalfund.org](http://www.theglobalfund.org)  
 Aeras, Global TB Vaccine Foundation: [www.aeras.org](http://www.aeras.org)  
 Drugs for Neglected Diseases Initiative: [www.dndi.org](http://www.dndi.org)  
 Global Alliance for TB: [www.tballiance.org](http://www.tballiance.org)  
 Institute for One World Health: [www.oneworldhealth.org](http://www.oneworldhealth.org)  
 International AIDS Vaccine Initiative: [www.iavi.org](http://www.iavi.org)  
 Malaria Vaccine Initiative: [www.maliavaccine.org](http://www.maliavaccine.org)  
 Access to this interactive links box is free online.

## Bring out your dead

**Despite recent progress, only a fraction of the drug industry's shelved compounds are shared with the research community. Could online collaborative research offer a solution?**

Failure—punctuated by the odd success—is an integral part of drug discovery and development. And industry is very vocal about the burden of its failed compounds in inflating the cost of bringing a drug to market—now purportedly \$2.56 billion according to the latest Tufts study. Industry is less forthcoming, however, when disclosing the thousands of ‘failed’ drug assets to the wider research community. In recent years, both the US National Center for Advancing Translational Sciences (NCATS) and the UK Medical Research Council (MRC) have made progress in coaxing de-prioritized compounds out of company vaults, with the latter announcing last month it had brought together “the world’s largest collection of de-prioritized compounds”—68 in all. But for drug repurposing of failed compounds to truly realize its potential, additional mechanisms need to be found that incentivize both large and small companies to release data so that they can be accessed and searched by the crowd.

Drug repurposing, or repositioning, is nothing new. Historically, it has often been a serendipitous process in which chance clinical observations suggest new indications for an approved drug or drug candidate. In recent years, compounds stalled in drug company pipelines due to lack of efficacy have received particular attention for repurposing. Such compounds are useful because they already have extensive information on safety and efficacy. Indeed, according to the Institute of Medicine, repurposed drugs can be approved faster (within 3–6 years of program initiation), at as little as ~60% the cost and with three times lower attrition rates than a drug from a traditional discovery program.

Two flagship initiatives that crowdsource researcher ideas for new uses are the NCATS’ Discovering New Therapeutic Uses for Existing Molecules program and the MRC’s Industry Asset Sharing Initiative. Key facets of either program are the provision of boilerplate collaborative research agreements, which streamline the legal and administrative burden between participating researchers and companies, and an open crowdsourcing application process followed by a closed second stage governed by cooperative agreements. Participants can apply for new use intellectual property (IP), with the companies donating the compounds getting first right of refusal.

Thus far, NCATS and MRC have succeeded in engaging eight companies from the entire industry. Although it is a promising start, it represents a tiny fraction of all discontinued programs.

Uptake has been slow for several reasons. First, the drug industry is conservative, particularly when repurposing initiatives might throw up safety signals about lucrative drugs in portfolios. Second, company insiders may dismiss crowdsourcing expertise—after all, they ‘know’ their compounds best (the NIH ‘not-invented-here’ problem). And third, sharing compounds and data for crowdsourcing has significant resource implications.

To work with NCATS or MRC, a company must rally all the relevant data on a discontinued compound; must locate internal champions with relevant expertise (who may be working on another asset or have left the

company); must appoint an individual to oversee the program (who would otherwise be working on another project in the organization); and, most importantly, must manufacture sufficient quantities of the compound at clinical grade to supply a trial. Industry insiders estimate these direct and ‘opportunity’ costs may total up to \$1 million per compound.

Clearly, this is prohibitive for many companies, especially the majority of small-to-medium-sized enterprises. Even if their overheads are less, most biotech companies have insufficient time, staffing and resources to participate in an NCATS/MRC collaboration. For companies that run out of funding, assets and IP are often left in limbo.

One alternative to the NCATS/MRC programs would be to create an online platform where users could upload data about shelved compounds in return for royalty options, milestone payments, and exclusive and non-exclusive licenses if compounds are taken forward. These online solutions could provide collaborative environments in which drug-related data would be uploaded, linking out to repositories like PubChem or ChEMBL. Researchers could either work in open environments to discuss ideas, organize research, and network or collaborate in private environments where proprietary information would be shared with a defined group.

The question is, who would build such a resource?

One thought is the European Union-funded Innovative Medicines Initiative (IMI). IMI has the authority, industry/academic contacts, and financial muscle to put the technology together in short order. Failing this, several internet-based collaborative platforms might also moonlight for this purpose. For example, Collaborative Drug Discovery’s (CDD) ‘Vault’ cloud-based offering to mine biological and chemical data has securely hosted over 250,000 user logins for 10 years now; Cures Within Reach is launching the CureAccelerator platform later this year, which has similar capabilities and the ability to network with funders.

The movement of drug discovery from silos in companies and other organizations to online platforms that enable collaborations with researchers across the globe is not going to happen overnight. But it is coming.

It will be enabled by the increasing ability to integrate molecular and structural data about compounds, data from electronic medical records and high-throughput large-scale phenotyping and genotyping technologies. It will be facilitated by greater transparency in disclosing clinical data on drugs, exemplified by the European Medicines Agency’s decision to publish from this month onwards complete clinical study reports for all marketing authorization applications. And it will be driven by the recognition that crowdsourcing the global research community can invigorate drug development with new ideas.

The upshot will be that collaborative drug discovery will become accessible to all companies, large or small, across the industry. If that means failures can be shared more widely, resources diverted from making the same mistakes and a greater number of failed compounds turned into successes, the sooner the better.

Corrected after print 8 January 2015.



## Erratum: Reinventing tech transfer

Brady Huggett

*Nat. Biotechnol.* 32, 1184–1191 (2014); published online 5 December 2014; corrected after print 9 December 2014

In the version of this article initially published online, the name of the University of Pennsylvania president Amy Gutmann was misspelled. The error has been corrected in the HTML and PDF versions of the article.

## Erratum: Bring out your dead

*Nat. Biotechnol.* 33, 1 (2015); published online 9 January 2015; corrected after print 9 January 2015

In the version of this article initially published, the cost of bringing a drug to market was incorrectly stated as \$2.3 billion. The correct amount, as per the Tufts study referenced in the article, is \$2.56 billion. The error has been corrected in the HTML and PDF versions of the article.

## Corrigendum: Linking T-cell receptor sequence to functional phenotype at the single-cell level

Arnold Han, Jacob Glanville, Leo Hansmann & Mark M Davis

*Nat. Biotechnol.* 32, 684–692 (2014); published online 22 June 2014; corrected after print 14 January 2015

In the version of this article initially published, the concentration of the V-region primers in the Online Methods section was given as 0.6  $\mu$ M. The correct concentration is 0.06  $\mu$ M. The error has been corrected in the HTML and PDF versions of the article.

## Corrigendum: Selling long life

Christopher Thomas Scott & Laura DeFrancesco

*Nat. Biotechnol.* 33, 31–40 (2015); published online 9 January 2015; corrected after print 14 January 2015

In the version of this article initially published, the caption for Figure 5 read “Pope Innocent VIII, likely the first patient to undergo parabiosis.” In fact, he did not undergo parabiosis, but a blood transfusion. The caption should have read “Pope Innocent VIII died in a rejuvenation attempt in 1492.” The errors have been corrected in the HTML and PDF versions of the article.

## Corrigendum: Status and market potential of transgenic biofortified crops

Hans De Steur, Dieter Blancquaert, Simon Strobbe, Willy Lambert, Xavier Gellynck & Dominique Van Der Straeten

*Nat. Biotechnol.* 33, 25–29 (2015); published online 9 January 2015; corrected after print 14 January 2015

In the version of this article initially published, Figure 3a had three errors. The heights for bars ‘26’ and ‘31’ for US, Vitamin E, broccoli + tomato + potato were ~7 and ~25, respectively; the transgenic (green) value US, vitamin E, tomato was given as ‘24.5’, but should be ‘40.3’. The errors have been corrected in the HTML and PDF versions of the article.

## Is open innovation the way forward for big pharma?

The current, fully integrated business model of large pharmaceutical companies is increasingly considered to be unsustainable, and so new approaches that engage large and small companies, governments and academic institutions are needed. Could 'open innovation' models that have proved successful in other sectors be fruitfully adopted by the pharmaceutical industry?

In the past, the business model adopted by most high-tech companies was traditional 'closed innovation'. Ideas were generated internally and taken from concept to commercialization using vertically integrated internal resources. Such closed organizations would guard all their intellectual property (IP) closely to protect company interests and to prevent exploitation by rivals.

Recently, however, there have been challenges to the efficiency of this model as technology has developed and the external environment has changed. As well as people moving more freely between organizations, taking their knowledge and expertise with them, the growth of the internet and the pervasiveness of media generally has meant that information is much more widely available. Consequently, for many R&D challenges, there is now a wealth of knowledgeable people distributed around the globe with the potential expertise to address them, which would be impossible for any individual company to hire. Tapping in to this expertise is one of the key goals of open innovation<sup>1</sup>.

### Examples of open innovation

Open innovation assumes a flexible business model in which new product innovation originates from both internal and external ideas. Opportunities to source expertise and potential products externally for further development, often through collaborations, are actively embraced, and internal ideas are allowed to be exploited outside the originating organization. Following on from pioneering demonstrations of open source development in the software industry — leading to products such as the Linux operating system<sup>2,3</sup> — other business sectors have adopted open innovation in varying ways. Although it is too early to say whether such strategies have been successful in most of these sectors, some advantages have already been demonstrated through different collaborative models.

For example, after the consumer products company Procter and Gamble (P&G) reshaped their R&D model to harness external innovation in 2000, it reportedly

increased its product success rate by 50% and the efficiency of R&D by 60%<sup>4</sup>. Another example is the mining company Goldcorp, which successfully utilized 'crowd sourcing' by creating a competition in which anyone could access its geological data online to identify potential new seams of gold, with prizes awarded to participants who submitted the best methods and estimates<sup>2</sup>. More than 8 million ounces of gold was found through the challenges, reducing expected exploration time by 2–3 years, and dramatically changing the fortune of the company.

### Open innovation in pharma

The pharmaceutical industry is facing many of the same, if not greater, challenges as the other sectors that have adopted open innovation approaches. Increased costs and risks; a need to understand better patient and payer requirements; and opportunities to gain access to better tools, technologies and ideas, have all driven large pharmaceutical companies to re-evaluate their business models. At the same time, academic biomedicine is realizing that there are tools, technologies and experience within pharmaceutical companies that can assist the development of their ideas and research towards clinical application. Consequently, there has been a shift in the cultural landscape that could allow a wider adoption of open innovation approaches.

Indeed, in addition to the rapid growth of partnering activity in the past decade, large pharmaceutical companies have already begun to work more actively together — as well as with small- and medium-sized enterprises and academic institutions — on pre-competitive research<sup>5</sup>. The success, in terms of numbers of projects funded and organizations involved, of the first call of the European Innovative Medicines Initiative (<http://www.imi.europa.eu>) is one example of this among the many other pre-competitive consortia globally.

Several large companies have also been experimenting with more radical open innovation models. For example, in 2001, Eli Lilly spun out InnoCentive as the first

Jackie Hunter is Senior Vice President of Science Environment Development at GlaxoSmithKline, Stevenage, Hertfordshire SG1 2NY, UK. Susie Stephens is Director of Biomedical Informatics at Johnson & Johnson Pharmaceutical Research & Development, 145 King of Prussia Road, Radnor, Pennsylvania 19087, USA. Correspondence to J.H. e-mail: [a.jacqueline.hunter@gsk.com](mailto:a.jacqueline.hunter@gsk.com) The views expressed herein are not necessarily those of the employer of S.S.

internet problem-solving platform designed to connect companies with research challenges to potential solution providers<sup>3</sup>. More recently, in 2009, Eli Lilly announced the Phenotypic Drug Discovery Initiative (<https://pd2.lilly.com/pd2Web/>), which makes their assays and expertise available to academic institutions to source new collaborations and compounds, and Pfizer began allowing other organizations to screen against their internal compound library. With the aim of harnessing large-scale biological data and tools, Sage Bionetworks (<http://sagebase.org/>) was launched in 2009 to build complex, predictive models of disease using an open innovation model, initiated by comprehensive data and analysis tools donated by Merck<sup>6</sup>. And in the field of neglected tropical diseases, GlaxoSmithKline (GSK) announced the creation of a patent pool, which aims to remove IP as a barrier to research into treatments for neglected diseases. GSK has put more than 800 patents for compounds or processes into this pool, and Alnylam has added a further 1,500 patents.

Support for open innovation models in general has recently been indicated by GSK's announcement that it is partnering to create a new bioscience park based on an open innovation model in which companies located on the park will have shared access to specialist skills, equipment and expertise. Johnson & Johnson's Head of Pharmaceutical Research & Development, Paul Stoffels, also announced a shift towards an open innovation model for the company in 2009.

### Misconceptions

Although such initiatives demonstrate that support for open innovation in the pharmaceutical industry is becoming increasingly widespread, one barrier to the adoption of such models has been the major misconception that open innovation is equivalent to open access, and in some way undervalues or undermines the concept of IP protection. However, IP is actually the currency of open innovation, and open innovation relies on the fact that creating solely owned and derived IP does not automatically lead to success and the creation of commercial value.

Indeed, across many industries, most patents remain uncommercialized; for example, Siemens and P&G recently reported that they only used 10% of their patent portfolio<sup>7</sup>. So, there is a real opportunity to create value from the sharing of IP. If this IP is not managed actively, the asset value is not realized. This becomes especially crucial in an environment in which the window of opportunity to recoup investment is short, as in pharmaceuticals. Proactive IP management looks to identify opportunities for sharing IP to create real value.

Such co-creation involves sharing the costs and benefits of innovation, and resulting IP, in line with the relative contribution of the various parties. This would include royalties and commercialization rights, but here IP rights are ensured through appropriate and harmonized protection strategies that are agreed on by all participants. At the academic-industrial interface, there has been some progress in the use of 'boilerplate agreements' (for example, the Lambert Agreements in the UK) that aid discussion starting from a point that

is appropriate to the particular collaborative situation (<http://www.innovation.gov.uk/lambertagreements/>). This is an approach that could be explored more widely for a range of sectors, including pharmaceuticals.

Another major misinterpretation of open innovation is that all collaborations require direct cash from industry to be of value, but this need not be the case. For example, many academic researchers are as motivated by the tangible and intangible benefits gained from partnering with industry, such as access to tools, reagents and expertise. The use of in-kind contributions can allow risk sharing and access to diverse thinking without upfront cash contributions, to the benefit of all parties, and such efforts deserve further focus to fully realize their potential.

### Moving forward

Adopting an open innovation ethos will require a change in the values placed by companies on new technologies and individual skill sets, as well as strong structural and cultural frameworks for optimal operation. New technology must be effectively deployed to maximize the benefits of open innovation; for example, the building of expertise networks and databases to allow the best partners to work with each other and dissemination of information about projects across an organization. Individuals should be employed who not only have relevant scientific expertise, but also possess strong external networks and are skilled in working with external organizations. Senior management must recognize that resources have to be applied to nurture collaborations and monitor their progress to ensure success. Employee roles may need to be redefined to ensure sufficient continuity with external organizations, and to ensure knowledge is being maintained within the enterprise.

Overall, in our view, open innovation is a valuable model for large pharmaceutical companies. It provides considerable flexibility, aiding the rapid ramp-up or -down of activities in particular areas, which would help the industry to keep up with the rapid pace of change in the external research community. Furthermore, the industry would then be better able to face the rapidly increasing cost of bringing products to market through risk-sharing deals. Finally, such a model would allow pharmaceutical companies to become closer to understanding patients and their requirements, and the needs of other stakeholders. This will not only be beneficial to the pharmaceutical industry, but also to academic institutions, biotechnology companies and, ultimately, to patients.

Jackie Hunter and Susie Stephens

1. Chesborough, H. W. *Open Innovation: the New Imperative for Creating and Profiting from Technology* 1–272 (Harvard Business School, Boston, Massachusetts, 2003).
2. Tapscott, D. & Williams, A. D. *Wikinomics: How Mass Collaboration Changes Everything* 1–320 (Portfolio, New York, 2006).
3. Munos, B. Can open-source R&D reinvigorate drug research? *Nature Rev. Drug Discov.* **5**, 723–729 (2006).
4. Huston, L. & Sakrab, N. Connect and Develop: Inside Procter & Gamble's new model for innovation. *Harv. Bus. Rev.* **84**, 58–66 (2006).
5. Barnes, M. R. *et al.* Lowering industry firewalls: pre-competitive informatics initiatives in drug discovery. *Nature Rev. Drug Discov.* **8**, 701–708 (2009).
6. Hughes, B. Harnessing open innovation. *Nature Rev. Drug Discov.* **8**, 344–345 (2009).
7. Alexy, O. *et al.* Does IP strategy have to cripple open innovation? *MIT Sloan Management Rev.* **51**, 73–77 (2009).



# Precompetitive consortia in biomedicine—how are we doing?

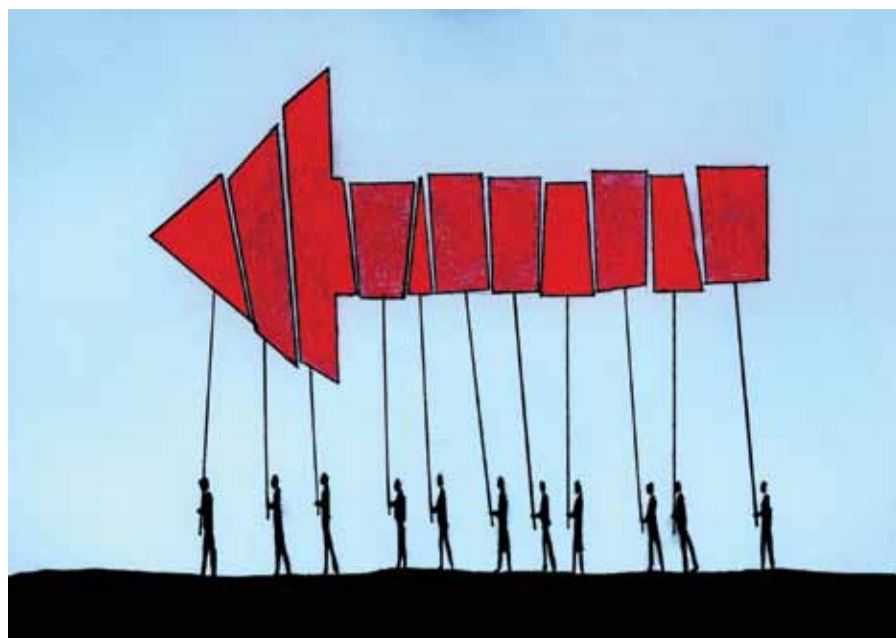
Barbara Mittleman, Garry Neil & Joel Cutcher-Gershenfeld

**Too few precompetitive consortia are being formed to mitigate lost opportunities and deliver on other potential mutual gains for public and private stakeholders in drug development.**

During the past decade, many drug companies entered the race to independently develop a small molecule inhibitor of p38 kinase for inflammatory disease. All reached the same conclusion—the target is not viable owing to fundamental toxicity problems related to its biology. Tens of millions of dollars were invested and countless hours of effort were expended for researchers to come to this realization. Yet it all could have been so different. Collaboration around the target through a precompetitive consortium would have decreased the real expenditures and opportunity costs while retaining the option for each company to develop proprietary therapeutics, had the target shown promise. Here we argue that despite the formation of consortia as a complement to market competition and government regulation in recent years, too few exist to mitigate lost opportunities and deliver on other potential mutual gains for public and private stakeholders in the drug development enterprise.

## What are consortia good for?

Consortia and cooperative arrangements, particularly precompetitive arrangements that generate shared resources such as data, tools and analytics, can help advance the drug development enterprise as costs increase, R&D



By pooling resources and expertise, precompetitive consortia enable progress in various drug discovery and development to move forward.

spending is cut back, regulatory requirements become more stringent, public trust in the industry declines and company valuations shrink<sup>1</sup>. At the core, too few new drugs and devices are making it to the market to alleviate the clinical need and to underwrite the spiraling costs of development. In fundamental ways, the processes, structures and cultures current in the industry are inadequate to meet the demands of biomedicine in the twenty-first century.

Shifting to a more sustainable drug development enterprise requires new models and methods for increasing efficiency and economy, decreasing and managing risk, stratify-

ing patients and developing new tools and approaches<sup>2</sup>. Broadly speaking, there are five strategic avenues for industry innovation: (i) independent operation of each stakeholder (a form of competition); (ii) mergers and acquisitions (M&As; a different form of competition); (iii) strategic alliances (a cooperative model that is similar to but distinct from precompetitive consortia); (iv) open innovation approaches, such as crowdsourcing, prizes and challenges (another cooperative model that is similar to and distinct from precompetitive consortia), and (v) precompetitive consortia (comprising, for example, industry, government, academia and public,

*Barbara Mittleman was at the US National Institutes of Health, Bethesda, Maryland, USA and is presently at Nodality, South San Francisco, California, USA; Garry Neil is at Johnson & Johnson, New Brunswick, New Jersey, USA; Joel Cutcher-Gershenfeld is at the University of Illinois at Urbana-Champaign, Illinois, USA.  
e-mail: [barbara.mittleman@nodality.com](mailto:barbara.mittleman@nodality.com)*

patient or professional advocates). Consortia offer unique opportunities for stakeholders to redefine the precompetitive space, develop new work streams and jointly produce tools and resources.

The synergies resulting from the consortium approach promise to save time and money by more efficiently and effectively generating therapeutics, diagnostics and devices to improve the public health. Although any avenue for innovation has the potential to deliver cost savings, cooperative models increase the potential for risk sharing. The consortium model best fosters collaboration among stakeholders with diverse interests and experiences. Consortia, combined with the existing competitive institutional and regulatory arrangements, constitute a more robust institutional landscape for the entire drug development enterprise.

Specific and demonstrable successes from biomedical consortia include the Biomarkers Consortium, the Predictive Safety and Toxicology Consortium, the Patient Reported Outcomes Consortium, the Genetics Association Information Network and the National Bone Health Alliance. These models exemplify collaboration in areas where individual companies were already active. But consortia can also encourage innovation. One such example of a cross-industry collaboration is TransCelerate BioPharma, which was formed not only to allow industry participants to collaborate more efficiently with one another but also to promote interaction among an 'ecosystem' of participants, including the US Food and Drug Administration (FDA), academia and patient advocates. Initial projects of this consortium include defining clinical data standards, developing common standards for investigator training and certification, creating a common portal for industry sponsored investigators and development of streamlined mechanisms to allow sponsors to obtain commercial comparator drugs for clinical trials<sup>3</sup>. This is not just an instance of precompetitive cooperation, but an explicit effort to better align the institutional context.

The success of precompetitive collaborations depends on a dynamic balance of cooperation, competition and regulation. At present, these cooperative aspects are underdeveloped, particularly in the United States. The use of this institutional arrangement is certainly more advanced in Europe. For example, the Innovative Medicines Initiative (IMI) represents a commitment of more than €2 billion (US\$2.7 billion) for multi-stakeholder collaboration. But this scale has been achieved through a structure in which industry has a strong role in driving the agenda—a setup that is helpful for generating industry contributions

but likely to be biased toward drug development and qualification rather than discovery and basic science.

So if consortia have such potential for benefit, why is this type of innovation not sweeping the industry? Objections to consortia are many and include claims that consortia act as sinks for time and effort, are slow and burdensome and reveal the differences in worldview and priorities across sectors without solving those differences, and that historical norms of secrecy are insurmountable. With this in mind, in the following sections we assess the return on investment for consortia and attempt to identify which types of value this approach is most likely to achieve and at what cost.

### The anatomy of consortia

The word 'consortium', derived from the Latin for fellowship, invokes joint goals, the synergy of shared action and the notion of an underlying (negotiated) agreement. At its core, a consortium involves a new form of alignment among independent but interdependent stakeholders. More precisely, we define the stakeholder alignment needed for a consortium to succeed as the extent to which interdependent stakeholders orient and connect with one another to advance their separate and shared interests.

Consortia provide mechanisms for stakeholders to 'orient and connect' with one another, and they deliver on both 'separate and shared interests'. Furthermore, stakeholders can be aligned (or misaligned) along distinct dimensions. Alignment can be behavioral (information sharing, trust and so on), structural (forums, incentives and so on), strategic (strategic intent and realized strategies) or cultural (underlying values and assumptions). In considering consortia, we consider all these dimensions of stakeholder alignment<sup>4</sup>.

Despite their underlying similarities of purpose, consortia differ markedly in the details of membership, operating policies and practices and the fates expected of the products of their activities. Which behavioral and structural differences are responsible for whether a consortium accomplishes its goals is an important and insufficiently understood question.

The effective use of precompetitive consortia depends on a nuanced understanding of their behavior, structure, strategies and culture, as well as the availability of appropriate instruments and methods to monitor and evaluate them. Developing these takes time, though probably less than the usual 12–15 years needed to take a new compound to market. The time function, therefore, has two aspects: the time required to establish and operate a consortium to completion, and the time required

to develop and apply assessment tools and to analyze the resulting data. An iterative cycle of assessment and improvement can then be instituted, optimizing the return on the investment of time and resources.

The development of a consortium requires the identification of a shared need and the complementary resources to be contributed by the participants, as well as the establishment of working rules that govern the shared activity and its products. The time required to do this is a function of factors, such as the specificity and concreteness of the task, the similarities or differences that animate the participants and the value of the products. The process of getting the Biomarkers Consortium to launch took approximately two years of negotiations, which focused on bridging the divergent standards and practices of industry, government and academia with respect to, for example, intellectual property. Although each project needs a tailored IP agreement, underlying consortium-wide policies provide a common point of departure for such agreements.

Coming to agreement regarding how the financing and resources will be arranged and overseen may be difficult. Most consortia in biomedicine depend on professional staff (which requires central funding) and additional resources to be targeted for selected projects. Aligning incentives—financial and nonfinancial—is key and requires an understanding of the reward systems in place for each of the participants. Consortium structures must ensure that each party receives the necessary individual benefits while accomplishing collective objectives, or the consortium is doomed to fail. Poorly aligned incentives cause failures, as they do not provide sufficient impetus to engage partners and elicit the commitment of necessary funds as competing priorities, opportunities and needs arise.

Likewise, investment in a consortium usually represents a very small fraction of the costs of a drug development program. Because the vast majority of drug development efforts fail in an expensive way—whether early or late—and the failure of a consortium is relatively inexpensive, experimentation in consortium development may well be cost effective. Cost effectiveness, in this context, relates to the potential value of the consortium process and products versus the value of doing it alone. Embedded in this notion of consortium value is the deeper notion of the contrast between proprietary and precompetitive aspects of the drug development value chain. Collaboration, in the framework of a consortium or other partnership, implies that sharing resources, risks and products can be mutually beneficial. In

## Box 1 Assessing the Biomarkers Consortium

After six years in operation, does the Biomarkers Consortium represent a good return on investment for its members? The consortium has launched 15 projects in areas such as Alzheimer's disease, cardiovascular disease and breast cancer, with several more projects now under consideration. The consortium's first project, 'The utility of adiponectin as a biomarker predictive of glycemic efficacy', was completed in 2009 (ref. 18); one other project has now been completed, and two more are slated for completion this year.

Through the consortium, industry has been able to constructively engage with the FDA on a precompetitive basis and understand the agency's thinking about biomarker qualification as a tool for drug development—opportunities that are useful for projects from the consortium as well as for other activities. At the same time, no qualified biomarker developed through the Biomarkers Consortium has yet delivered tangible results to patients or other stakeholders. Is this what must happen if continued engagement is to be expected in the consortium?

contrast with a joint venture or strategic alliance, a precompetitive consortium involves more open sharing, a risk that is balanced against the potential benefit of inputs from multiple stakeholders. At the other end of the cooperative spectrum, open innovation, such as crowdsourcing, involves more transparency and proportionally greater risk, counterbalanced by the potential for more diverse and unexpected inputs. In these senses, precompetitive consortia represent a middle path between risk and benefit.

Precompetitive sharing of tools and platforms can facilitate comparison across industry and academia, facilitate regulators' abilities to make data-driven assessments and clarify regulatory principles without impinging on the business model of big pharma<sup>5</sup>. Similarly, some argue that target sharing should also fall into the precompetitive space, allowing cost and risk sharing early in the development process without interfering with later-phase proprietary development of leads, as is currently done<sup>6</sup>. As noted above, p38 kinase target programs initiated by many companies hit a dead end that represented a massive loss of resources to biomedicine as a whole—in the direct costs as well as in opportunity costs from other, more viable programs not being pursued.

As demonstrated by the Biomarkers Consortium (see **Box 1**), however, managing expectations, understanding the real potential and limitations of consortia and carefully monitoring and evaluating progress and outcomes are key to understanding the place of consortia in the complex ecosystem of drug development. Ultimately, the fundamental question for all stakeholders in drug development is whether the continued, current independent patterns of interaction will be sufficient to sustain the enterprise in the future, or whether a rebalancing of competition and collaboration is needed.

### Surmounting drug development's drawbacks

The inefficiencies and high attrition rates of current drug development are well known<sup>7,8</sup>. These challenges arise on several levels.

**Structural and strategic challenges.** Not only is the overall process for drug development in crisis, but there is also ambiguity in the overall strategy. Niche drugs with stable but limited markets are often discarded in favor of potential blockbusters targeting common conditions. Focusing on disease entities rather than underlying pathobiological pathways dictates clinical trial designs that fail to account for heterogeneity among patients. The lack of biomarkers and other patient-stratification tools prevents optimization of trials to avoid toxicity and maximize efficacy. Regulatory decision making would be facilitated and improved by the availability of data and qualified tools<sup>9</sup>. The barriers arise from limitations in understanding of the biology, validated targets and adequate translational tools. Such understanding does not provide competitive advantages, and failing to share it creates major practical disadvantages.

Pooling of resources, data, tools and platforms—and sharing of the resulting products—offers a way to streamline drug development. Consortium arrangements allow the harmonization of tools and markers developed through shared efforts and shared in a precompetitive manner, thereby providing new standards and regulatory benchmarks that are informed by science, technically validated and open for qualification for use. Consortia also offer the opportunity to explore new clinical trial approaches such as to focus efforts through adaptive designs, stratify patients according to risk or markers, test across disease boundaries according to implicated molecular pathways and/or interrogate data sets for new hypotheses and signals. Thus, consortia offer opportunities for resource leverage, precompetitive assay

development, knowledge sharing, target validation and so on. Such sharing, in turn, enables and promotes competition by allowing more and higher-quality molecules to be developed by companies that will compete using these new drugs in the marketplace.

The paradox that precompetitive sharing promotes, rather than impedes, competition depends on carefully defining the crux of the competitive business model for the industry<sup>10</sup>. Pharmaceutical companies generate value and income through the development and marketing of medicines. Cost centers that siphon off resources reduce overall profits, and reductions in such costs improve the return to investors. Identifying cost centers that can be reduced by cost sharing provides value and increases return on investment.

Designing such cost-saving and risk-sharing arrangements depends on engaging the appropriate partners in consortia. From a structural standpoint, the drug development ecosystem can be said to comprise six major types of stakeholders: (i) pharma; (ii) biotech (mainly small and medium-sized companies (SMEs)); (iii) academia; (iv) government labs (such as the US National Institutes of Health (NIH)); (v) government regulators (such as the FDA), and (vi) patients, non-profit organizations and civil society. Each stakeholder is driven by distinct motives and incentives, answers to distinct authorities and has individual metrics for progress and success. Common to all, however, is the goal of improving health and well-being for patients. Major changes, at increasing pace, are occurring within each stakeholder community, and these changes increase instability in the ecosystem and provide pressure and opportunity for cross-sector collaborations. New forms of behavioral, structural and cultural alignment—in which adjustments are made by all stakeholders, taking into account their separate and shared interests—will be needed.

The dynamic structural landscape is characterized by M&As that have reduced the number of large pharma companies by nearly half. Furthermore, cutbacks in pharmaceutical R&D funding, dissolution of discovery groups and shuttering of facilities are common following M&As. Many small biotech companies have changed their business models, shifting from fully integrated pharmaceutical companies to simply targets for purchase by a larger company (preferably a pharmaceutical company) through advancing an area, target or molecule to the point where the larger organization can take it through the clinic (thereby substituting the smaller company for in-house R&D capabilities). Thus, new drugs originate increasingly, albeit inefficiently and unpredictably,



in smaller enterprises. The lack of efficiency and predictability is manifest as an inadequate return on investment and provides justification for the reduction in R&D both within companies and after M&As. Academic stakeholders are also in a state of flux, with public universities facing funding crises and organizational challenges to facilitating interdisciplinary work. Traditional tenure and promotion criteria do not fully recognize the value of team science and of crossing departmental boundaries. Finally, government labs and regulatory authorities face the cross-currents of increasingly partisan politics and fundamental debates over the level and nature of government spending on many aspects of health.

Neither universities nor government is configured to develop, test and market medicines, and it is not clear that these activities are consistent with the mission of either. The academic mission to educate and to generate knowledge has important contributions to make to drug development, as do the scientific and regulatory missions of government agencies, but in neither case does that directly involve the business of medicines. Patient and public advocacy organizations and other mission-driven nonprofits are playing an important part in filling gaps not fully served by market forces, but this is an idiosyncratic process that depends on the

emergence of leadership in a given disease area. Despite the emergence of venture philanthropy, patients are poorly represented in the drug development decision-making process overall. A further limiting factor is the availability of patients to enroll in clinical trials and to serve as a source of specimens and data. This is particularly true in circumstances where a disease often occurs with independent but co-morbid conditions, causing numerous exclusions from entry into clinical trials, and/or where patient populations are small.

The present misalignment of organizational or sector goals and incentives, however, constrains the ability of the ecosystem to accelerate the delivery of cost-effective solutions for health problems. These are the opportunities where consortia can make a difference.

**Cultural challenges.** Structural and strategic challenges that make alignment of parties and incentives difficult are also accompanied by deep cultural challenges<sup>11</sup>. Secrecy and privacy as operating principles are deeply ingrained in the pharmaceutical industry—is designed to protect market advantage by hiding the identity of targets of interest, clinic and therapeutic areas under investigation, and even sharing between preclinical and clinical teams within the company is unusual in many organizations.

There are, of course, major cultural differences among and within companies, owing in part to the differences in nature between small, startup organizations and large, venerable, complex and isolated organizations with established ways of acting.

Of course, industry is not the only sector that is culturally diverse. Within government, cultural differences exist between the FDA, the NIH and the Center for Medicare and Medicaid Services, as a result of the agencies' diverse missions as well as the training, experience and habits of the people working there<sup>12</sup>. Within academia, there are cultural differences between institutions regarding attitudes toward technology transfer and the relative values of translational research and basic science.

Perhaps the deepest cultural challenge involves the lack of understanding in the general population of the risk that accompanies all therapies. Finally, important distinctions and deeply rooted cultural assumptions about health and disease are in the background of each sector and individual within the system.

Having drawn these broad characterizations, it is also clear that no sector or organization is monolithic; distinct (micro-) cultures exist within many or all companies, agencies and organizations. Aligning incentives, promoting effective and substantive communication and developing viable shared practices depends on a sensitivity to the cultural bases of behavior, expectations and willingness to assume risk or uncertainty.

### A proposal for wider adoption of consortia

We propose that, if cross-sector consortia are used judiciously, drug development writ large can more effectively promote the broad needs of patients and society, competition and the profit motive can drive innovation in the marketplace, regulation will enforce public health and safety standards, precompetitive collaboration will reduce waste and minimize risk, and collaborative approaches to science will open up new frontiers. Elements of this vision currently motivate the drug development enterprise at all levels. Missing, however, are mechanisms that align and orient the stakeholders so that it is possible for them to function together in a coherent and efficient system. Ad hoc approaches to drive change have been undertaken by subsets of stakeholders, including the NIH Roadmap, the FDA Critical Path Initiative and the European Commission's IMI. Academic and commercially sponsored conferences addressing crucial issues—such as the intersection of disease and wellness, the handling of 'small *n*' diseases, the boundary between

## Box 2 Elements of effective, multi-stakeholder consortia

Although there is no one-size-fits-all solution, several common elements are shared by successful and effective consortia. These are:

- **Shared goals and vision.** A clear statement of the consortium's goals and overall vision for success (including a shared understanding of what these words mean).
- **Engaged stakeholders.** A definition of a governance structure and rules for decision-making.
- **Leadership and trust.** Leaders able to be effective on the basis on influence, rather than authority, with trust among principals providing a necessary foundation as policies and norms develop.
- **Roles, responsibilities and rights.** Well-defined roles, responsibilities and rights for all participants—typically in the form of a charter, with monitoring, learning and conflict-resolution mechanisms that enable necessary adjustments over time.
- **Policies, protocols and standards.** Well-articulated policies, protocols and standards governing the workings of the consortium and how the participants relate to one another (for example, in the areas of confidentiality, conflicts of interest, intellectual property, consortium grants and contracts, data access and data sharing, finances and flow of resources).
- **Milestones and benchmarks.** Benchmarking and communication processes to ensure that all parties—participants, stakeholders and beneficiaries, including regulators, clinicians, patient groups and advocates, the public, and so on—have an appropriate understanding of the risks and accomplishments.
- **Exit and transformation plan.** A built-in capacity for transformation, should fundamental changes in strategy, structure and process be necessary. An 'exit plan' should describe the conclusion of the consortium either when goals have been met or when it is determined that goals, milestones or benchmarks cannot be met.

the precompetitive and the competitive and the best processes for road mapping—have been mounted, but only by subsets of stakeholders.

A skillful mix of collaboration, competition and regulation based on both lateral and cross-domain alignment of stakeholders can lead to the efficient and effective development of consortia, which, coupled with well-managed and reasonable expectations, could define the rightful place of this approach within the repertoire of drug development tools and approaches. Optimizing consortium structure and ensuring its 'goodness-of-fit' to the activities and goals of the consortium are crucial elements of success.

No single structure or set of terms is universally useful<sup>13</sup>. However, all successful and effective consortia, share the following components (see **Box 2**): (i) well-defined shared goals and objectives, (ii) well-specified governance and decision-making responsibilities, (iii) clearly articulated roles, responsibilities, contributions and returns of the partners, (iv) appropriate policies defining, for example, antitrust, publication and IP that are spelled out at the outset and adapted through experience, (v) leadership based on influence more than authority, and (vi) mechanisms for shared learning and continuous improvement. Further policies may also be needed and should be customized to the specific partnership or consortium. Finally, sufficiently specific and well-defined milestones and benchmarks permit concise understanding of a consortium's 'end points' and allow allocation of the correct amount of capital for a dedicated problem<sup>14</sup>. Once capital allocation and milestones are defined, risk is lowered and expectation of success is inherently more realistic. The advantages of consortia must therefore be understood from two distinct vantage points: that of strategy and that of realizing operational efficiencies.

Criticisms of consortia are generally based on the long time required for them to generate results, their excessive bureaucracy, the risks in sharing information, and specific regulatory barriers around certain forms of collaboration. But pooling resources has been shown to hasten the availability of data and tools, resulting in the availability of better biomarkers, improved clinical trial designs, expanded data sets and other advances. Likewise, the ability of regulators to access consensually validated and broadly available tools is valuable to facilitate effective regulatory decision making. The decision to develop a consortium rests on balancing the strategic and operational benefits to be gained the risks in time and resources and the likelihood of success. The goals and the operational structures developed to accomplish them can be translated into benchmarks

and metrics. Monitoring the progress of the consortium toward its goals can also permit the timely dissolution of arrangements that are not succeeding, allow necessary course corrections and point to improvements in structure and/or functioning of the partnership.

The subject of when and how to use the consortium approach and how to assess its success is itself ripe for scientific investigation<sup>15</sup>. Such a 'science-of-science' approach is founded on underlying theory-based principles, including the notion of internal and lateral alignment of parties in negotiation with one another<sup>16,17</sup>. Major pharma companies are complex and internally diverse organizations—as are government agencies and universities—and internal alignment and even internal communication can be very challenging and not culturally expected. Coordination within companies to determine when it is better to operate independently and when to work in partnership with others is relatively new. Some major pharmaceutical companies, such as Johnson and Johnson, now have internal forums in which company representatives involved in different consortia compare notes and develop coordinated strategies for engagement with consortia<sup>14</sup>.

The institutional arrangements associated with consortia are still in the early stages of development compared with the infrastructure associated with M&As. Consortia represent a contrast with the established systems for identifying and pursuing acquisition targets as a source of innovation. Consortia effectively round up all the potential buyers (pharma) of a technology and can give them access to the technology at 'bargain-basement' rates, with some development funding as the price of admission. The result is that what could be a market-driven, inter-company, competitive bidding process for a technology asset may become, through a consortium, a one-time sale to a pharma buyers' club spanning much of the market. A small company in a consortium risks having its IP accessed by all its potential customers but gains an inside track for visibility. And the risks of not participating may be worse—the consortium might fund an alternative route or breakthrough. It is possible that in the future, investors may even hold it against companies that are not in on the ground floor with consortia that are in an appropriate strategic space, just as they now will hold it against a firm (large or small) that is not managing the M&A space well.

Consortia are not yet well institutionalized across the industry or among other stakeholders. By virtue of their diverse membership, divergent underlying cultures and varying business structures and practices, they are dynamic

arrangements requiring alignment within each stakeholder organization as well as among the stakeholder organizations themselves. When linear assumptions of cause and effect are applied to complex systems, the outcomes may be unanticipated, especially if little attention has been given to mechanisms for feedback and adjustment. Clear structures and policies for shared activities, as well as defined benchmarks and strategies for change, provide a framework for consortium success and for understanding the forces that govern that success.

Achieving lateral alignment (among stakeholders) is a process that may be quite unfamiliar to parties accustomed to working within hierarchical structures. Consortia are characterized by lateral and multi-layered relationships, requiring more influence and problem solving than authority and direction—skills that can vary considerably across consortium participants. Clarity in structure and shared vision regarding the goals and manner of interaction among the participants serve to provide common working rules and to promote sustained alignment across stakeholder groups, thereby increasing the ability of the consortium to accomplish its goals and to satisfy its participants. Maintaining consistent vision and goals, trust relationships and leadership can be challenging given the relatively high rate of personnel turnover or change of roles within industry, as well as in other stakeholder sectors.

### Challenges of consortia

Better metrics and measurement are needed to assess whether biomedical consortia are progressing toward their goals and delivering value to their participants. Individual stakeholders will pull back if they do not see their interests advanced, yet if they are focused only on self-interest, the consortium will not be sustainable. Thus consortia need to deliver on a mix of individual and collective interests among diverse stakeholders. **Table 1** presents an initial schema of parameters and metrics that can assist in the monitoring of a consortium and help to provide guidance for course corrections and/or disbanding the consortium when goals are met or when it becomes apparent that they will not be met.

There are challenges apart from those of establishing the needed constituents, resources, structures and operations and of setting up the appropriate monitoring systems to ensure timely accomplishment of benchmarks, milestones and metrics, but they are not as easily quantified, assessed or rectified. Internal alignment within member organizations occurs outside of the view of the consortium but is crucial to the joint venture's success<sup>13,18</sup>. It is important that

**Table 1** Benefits, risks and relevant metrics for consortia

Goals and objectives	Benefits	Risks	Sample protocols and metrics
Prevent redundancy and waste	<ul style="list-style-type: none"> <li>Reduction of cost</li> <li>Enabling of diverse inputs</li> <li>Mitigation of risk of parallel dead ends</li> <li>Ability to 'fail early, fail cheap'</li> </ul>	<ul style="list-style-type: none"> <li>Expenditure of time and effort to reach agreement or consensus</li> <li>Unwieldiness and time costs of bureaucracy and operational processes</li> <li>Differences among stakeholders in willingness and timing to kill projects</li> </ul>	<ul style="list-style-type: none"> <li>Assess stakeholder alignment, both internal and external</li> <li>Determine lead time to initiation of activities</li> <li>Assess progress by setting milestones</li> <li>Model opportunity costs of participation versus those of nonparticipation</li> </ul>
Provide access and synergy	<ul style="list-style-type: none"> <li>Access to new tools, samples, platforms, patients and so on</li> <li>Facilitation of work that could not be done by any one stakeholder alone</li> <li>Identification of combined opportunities as a result of diverse interests</li> </ul>	<ul style="list-style-type: none"> <li>Increase in risk of competitors gaining insight into internal strategies and plans</li> <li>Driving of increased sharing of resources and information by expectations of reciprocity</li> </ul>	<ul style="list-style-type: none"> <li>Measure access and cost of access</li> <li>Model strategic risk/benefit ratio</li> <li>Establish mechanisms for 'modular' participation</li> </ul>
Shared cost	<ul style="list-style-type: none"> <li>Reduction of front-end expenditures</li> <li>Full benefit at fractional cost</li> </ul>	<ul style="list-style-type: none"> <li>Cost of participation in personnel costs and other infrastructure costs</li> <li>Reduction of value owing to shared benefit</li> </ul>	<ul style="list-style-type: none"> <li>Measure actual costs</li> <li>Model benefit/cost ratio</li> </ul>
Shared risk	<ul style="list-style-type: none"> <li>Shared responsibility</li> <li>Opportunity to influence risk profile</li> </ul>	<ul style="list-style-type: none"> <li>Diffuse credit</li> <li>Unpredictable sharing of blame</li> </ul>	<ul style="list-style-type: none"> <li>Model risk/benefit ratio</li> </ul>
Networking, contact and information exchange	<ul style="list-style-type: none"> <li>Engagement of new parties in the work</li> <li>Acquisition of useful information and insights</li> <li>Chance to convey useful perspectives and influence others</li> <li>Reduction of traditional antitrust liability risk</li> </ul>	<ul style="list-style-type: none"> <li>New form of antitrust liability possible</li> <li>Confidentiality risks</li> <li>Changes in market share/prospects</li> <li>Information exchange that benefits some stakeholders more than others</li> </ul>	<ul style="list-style-type: none"> <li>Assess and/or model market prospects</li> <li>Assess and/or model stakeholder returns on investment</li> <li>Assess changes in regulatory policy, market forces, public opinion and so on.</li> </ul>
Transparency	<ul style="list-style-type: none"> <li>Improved public trust</li> <li>Ease of joining and participating</li> <li>Knowledge of risks, benefits, outcomes, milestones and metrics</li> <li>Greater visibility of perspectives of less powerful stakeholders (particularly under consensus decision-making rules)</li> </ul>	<ul style="list-style-type: none"> <li>Reduced exclusivity</li> <li>Decreased flexibility and agility as a result of increased structure and operational predictability</li> </ul>	<ul style="list-style-type: none"> <li>Measure impact within stakeholder groups</li> <li>Measure 'organizational drag' imposed by consortium rules and operations</li> </ul>

members assign representatives who have sufficient gravitas within their own organizations and the authority to commit the organization to the consortium. These representatives face predictable dilemmas—they may be accused in their own organizations of having 'forgotten where they came from' and having 'gone native' while being accused in the consortium of not moving fast enough to ensure progress.

Trust among the members is key—particularly at pivotal moments in the life of the consortium, such as its formation—but difficult to force and measure. Without fundamental trust, commitment to persevere through points of disagreement and misaligned goals can be tentative or insufficient. Change in member personnel is an example of a pivotal event in the operation of a consortium with potential to undermine continuity, trust, efficiency and shared vision. Changes of personnel occur frequently in industry and other sectors, so this is a predictable challenge. Consortium policies and practices must be written to accommodate communication between leaders and members and among members as a means not only to exchange information but also to embody the

commitment to sharing, trust, transparency and maintaining a common vision.

Precompetitive sharing should provide benefit to all of the participants and often to the public and nonparticipants as well. The motivation and ability of a stakeholder to join such an effort depends on several factors: whether consortium goals fall outside of or are tangential to the central business model or revenue center of the organization; whether the organization can tolerate devoting manpower, energy and possibly monetary and/or tangible resources to the effort; and whether the potential benefits of sharing exceed the risks. Of course, not all stakeholders can engage equally in precompetitive consortia.

Biomarkers, for instance, are precompetitive tools for the pharmaceutical industry, whose profit model centers around the registration and marketing of therapeutics. For a small biotech or device company, however, the discovery, development, qualification and marketing of biomarker tests is not precompetitive and may be the center of their financial model. Likewise, information sharing may prove much riskier for an SME with a limited portfolio of IP and know-how. On the other hand, consortia

offer smaller enterprises a greater chance that their perspectives will be seen and valued by regulators and other leaders. Transparency and open sharing of information can markedly alter the usual licensing and sales process, either undermining or enhancing the ability of the SME seller to market itself and its product(s). Protecting the competitive edge and promoting sharing as a means to accomplish efficiencies of scale and synergies from bringing diverse resources together are counterbalanced, then, by the needs of academics to maintain their exclusive roles as experts, inventors and discoverers of knowledge; of SMEs to thrive in the market ecosystem (by bringing a product to market or by selling themselves as a product in the market); and of large corporations and SMEs to sustain a tenable business plan.

### Concluding remarks

Factors challenging drug development, such as rising costs, increasing failure rates and regulatory uncertainty—set in a climate of global economic strain and increasing pressures on the costs of health care—are pushing the pharmaceutical industry to become leaner, more efficient and more effective in delivering



products to the public. The present system is not working. Among available options, precompetitive consortia, with defined goals and structures that align the incentives of all stakeholders, provide an option that can be strategic and low cost to individual stakeholders and can facilitate the delivery of shared tools and approaches to reduce both cost and risk.

Ultimately, any system has four basic elements: inputs, processes, outputs and feedback. In the drug development ecosystem, there is substantial variability in all these elements. The building and examining of the role of consortia is an opportunity for a rebalancing of the competitive, collaborative and regulatory aspects of the system. Competition should be sustained in domains where it continues to drive innovation and serve society. In those where efforts are being duplicated, and resources are being wasted, precompetitive consortia and other mechanisms may produce better outcomes for the respective stakeholders and the system as whole. Consortia provide regulators and regulated parties more opportunities for mutual learning, which serves the overall public interest.

Consortia are just one tool to address systemic problems in biomedicine. Here, we

have presented a proposal for how to increase the utilization and effectiveness of consortia. In fact, the biomedical enterprise will ultimately need to be realigned to achieve a more dynamic mix of competition, collaboration and regulation if it is to continue to deliver on the promise of improved societal health through the acceleration of change in science, technology and society.

#### COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

#### ACKNOWLEDGMENTS

We appreciate support for this research provided under NSF-VOSS EAGER 0956472, "Stakeholder Alignment in Socio-Technical Systems."

1. FDA. The beginnings: laboratory and animal studies. *Information for Consumers (Drugs)* <http://www.fda.gov/Drugs/ResourcesForYou/Consumers/ucm143475.htm> (2011).
2. Miller, G. Necessity driving leery pharma "open." *FiercebiotechIT* <http://www.fiercebiotechit.com/story/necessity-driving-leery-pharma-open/2011-04-04> (2013).
3. Pollack, A. Drug makers join efforts in research. *New York Times* (20 September 2012).
4. Cutcher-Gershenfeld, J. & Lawson, C. Valuing the commons: a fundamental challenge across complex systems. (NSF, Washington, DC, 2010).
5. Bunin, B.A. & Ekins, S. *Drug Discov. Today* **16**, 643–645 (2011).
6. Lessl, M., Bryans, J.S., Richards, D. & Asadullah, K. *Nat. Rev. Drug Discov.* **10**, 241–242 (2011).
7. DiMasi, J.A., Hansen, R.W. & Grabowski, H.G. *J. Health Econ.* **22**, 151–185 (2003).
8. Toronto International Data Release Workshop Authors. *Nature* **461**, 168–170 (2009).
9. Hunter, J. & Stephens, S. *Nat. Rev. Drug Discov.* **9**, 87–88 (2010).
10. Munos, B. *Clin. Pharmacol. Ther.* **87**, 534–536 (2010).
11. Waitz, I., Townsend, J., Cutcher-Gershenfeld, J., Greitzer, E. & Kerrebrock, J. Aviation and the environment: a national vision statement, goals and recommended actions. [http://web.mit.edu/aeroastro/partner/reports/congrept\\_aviation\\_envirn.pdf](http://web.mit.edu/aeroastro/partner/reports/congrept_aviation_envirn.pdf) (FAA/NASA, Washington, DC, 2004).
12. Masum, H. & Harris, R. Open source for neglected diseases: magic bullet or mirage? (Results for Development Institute, Washington, DC, 2011).
13. Cutcher-Gershenfeld, J., Mittleman, B., Dickherber, A., Mayrand-Chung, S. & Franks, A. Stakeholder alignment in the Biomarkers Consortium. (NSF-VOSS EAGER 0956472, Washington, DC, 2012).
14. Woelfle, M., Olliaro, P. & Todd, M.H. *Nat. Chem.* **3**, 745–748 (2011).
15. Cutcher-Gershenfeld, J., Barrett, B. & Lawson, C. Building the internal organization to support lateral alignment: a case study of the Office of Environment and Energy, Federal Aviation Administration <http://ssrn.com/abstract=2345248> (MIT, Cambridge, Massachusetts, 2005).
16. Walton, R. & McKersie, R. *A Behavioral Theory of Labor Negotiations*. (McGraw Hill, New York, 1965).
17. DiMasi, J.A., Hansen, R.W. & Grabowski, H.G. The price of innovation: new estimates of drug development costs. *J. Health Econ.* **22**, 151–185 (2003).
18. Wagner, J.A. et al. *Clin. Pharmacol. Ther.* **86**, 619–625 (2009).